



Towards responsible development and application of large language models

Emily M. Bender
University of Washington

HPLT & NLPL Winter School
7 February 2023



@emilymbender
@emilymbender@dair-community.social

Outline



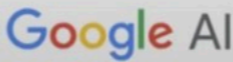


- Problematization of the rush for scale and the “foundation models” conceptualization
- Risks associated with large (and ever larger) LLMs
- Evaluation & its prerequisites
- Value sensitive design and techniques for mitigating risk
- Dangers and responsibilities that come with working on a ‘hot topic’

...

Great slide from [@chrmanning](#) showing the crazy rate of NLP progress.

This is just the beginning.

2018 NLP breakthrough with big language models

ELMo, ULMfit Jan 2018 Training: 103M words 1 GPU day	GPT June 2018 Training 800M words 240 GPU days	BERT Oct 2018 Training 3.3B words 256 TPU days ~320-560 GPU days	GPT-2 Feb 2019 Training 40B words ~2048 TPU v3 days according to a reddit thread	XL-Net, AIBERT, Grover, Megatron-LM, T5, ERNIE, RoBERTa, ELECTRA, GPT-3 , LaMDA, PaLM, Mistral, Wenxin, Gopher, ... July 2019-now
				

11:44 AM · Dec 8, 2022

GPU go brrrr

- ... and the number goes up.
- But which number?
 - Amount of text in the training set
 - Compute time
 - Benchmark scores (see Raji et al 2021)



Jack Rae ✓
@drjwrae



A new episode of the “bitter lesson”: almost none of the research from ~2 decades of dialogue publications, conferences and workshops lead to #ChatGPT. Slot filling ✗ intent modeling ✗ sentiment detection ✗ hybrid symbolic approaches (KGs) ✗

3:42 AM · Dec 9, 2022



@emilybender@dair-community.social on Mastodon
@emilybender



The bitter lesson is how much of the field is willing to accept a system that produces form that *looks like* a reliable solution to task as actually doing the task in a way that is interesting and/or reliable.

Are we doing science or just standing in awe of scale?



Jack Rae ✓ @drjwrae · Dec 9, 2022

A new episode of the “bitter lesson”: almost none of the research from ~2 decades of dialogue publications, conferences and workshops lead to #ChatGPT. Slot filling ✗ intent modeling ✗ sentiment detection ✗ hybrid symbolic approaches (KGs) ✗

2:59 PM · Dec 9, 2022

From yesterday's blog post from Pichai of Google

Since then we've continued to make investments in AI across the board, and Google AI and DeepMind are advancing the state of the art. Today, the scale of the largest AI computations is doubling every six months, far outpacing Moore's Law. At the same time, advanced generative AI and large language models are capturing the imaginations of people around the world. In fact, our Transformer research project and our field-defining paper in 2017, as well as our important advances in diffusion models, are now the basis of many of the generative AI applications you're starting to see today.



@emilybender@dair-community.social on Mastodon
@emilybender




More tales from the front of [#NLProc](#)'s evaluation crisis. What is PubMedGPT actually for and why are medical licensing exam questions a legitimate test of its functionality in that task?

>>



Percy Liang @percyliang · Dec 15, 2022

 CRFM announces PubMedGPT, a new 2.7B language model that achieves a new SOTA on the US medical licensing exam. The recipe is simple: a standard Transformer trained from scratch on PubMed (from The Pile) using @mosaicml on the MosaicML Cloud, then fine-tuned for the QA task.

[Show this thread](#)

3:57 PM · Dec 16, 2022 · **14.3K** Views



@emilybender@dair-community.social on Mastodon

@emilybender



From further down [@percyliang](#)'s thread, apparently it's not really "for" anything yet. All of that is "future work".

This is the core problem with the "foundation models" conceptualization. They are impossible to evaluate.

[twitter.com/percyliang/sta...](https://twitter.com/percyliang/status/1598888888888888888)

>>



Percy Liang @percyliang · Dec 15, 2022

We hope that PubMedGPT can serve as a foundation model for biomedical researchers; can it be adapted fruitfully for tasks such as medical text simplification, information retrieval, and knowledge completion? There's a lot more to do!

[Show this thread](#)

3:59 PM · Dec 16, 2022 · **6,531** Views

“Against Scale: Provocations and Resistances to Scale Thinking” (Hanna & Park 2020)

- “Scale thinking” prioritizes scalability, which in turn requires interchangeability of components (including workers) and users
- Hanna & Park ask:
 - Does the technological system centralize power (either through coordination, data extraction, or authority) or distribute it between developers and users?
 - Does the technological system treat the contributions and experiences of individuals as interchangeable or as uniquely essential?
 - Does it open up avenues for participation, and are those avenues of participation mobilizing or demobilizing?

“Against Scale: Provocations and Resistances to Scale Thinking” (Hanna & Park 2020)

- “Scale thinking” prioritizes scalability, which in turn requires interchangeability of components (including workers) and users
- Let’s ask, regarding LLMs:
 - Does the technological system centralize power (either through coordination, data extraction, or authority) or distribute it between developers and users?
 - Does the technological system treat the contributions and experiences of individuals as interchangeable or as uniquely essential?
 - Does it open up avenues for participation, and are those avenues of participation mobilizing or demobilizing?

Outline

- Problematization of the rush for scale and the “foundation models” conceptualization
- Risks associated with large (and ever larger) LLMs
- Evaluation & its prerequisites
- Value sensitive design and techniques for mitigating risk
- Dangers and responsibilities that come with working on a ‘hot topic’

- 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, 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?



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 (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
- Black Lives Matter 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



What are the risks?

Research trajectories

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

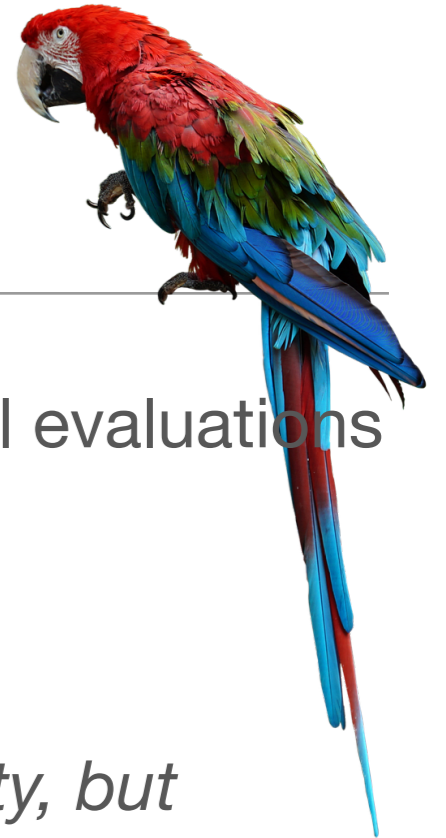


- These harms largely stem from the interaction of the ersatz fluency of today's language models + human tendency to attribute meaning to text
- Deeply connected to issue of accountability:
 - Synthetic text can enter conversations without anyone being accountable for it
- Accountability key to responsibility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



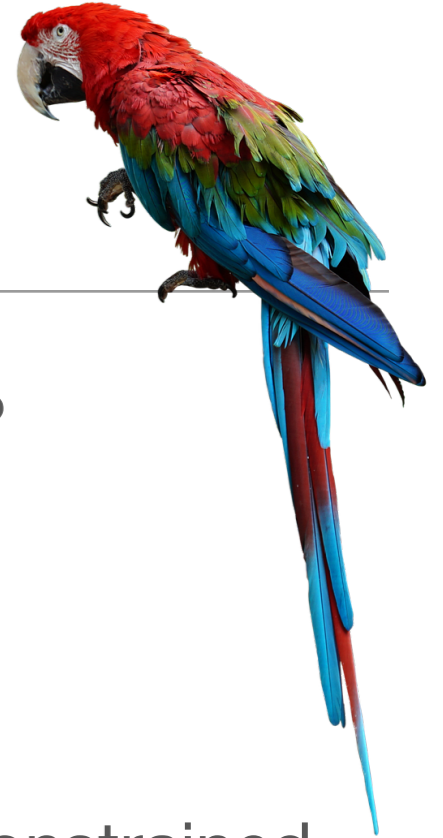
Risk management strategies

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?

The view from 2023

- 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 2023

- 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 2023

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

Helping you when there's no answer

MUM has the potential to transform how Google handles search tasks. MUM uses the [T5 text-to-text framework](#), which is 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/>

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 2023

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

Helping you when there's no answer

MUM has the potential to transform how Google handles search tasks. MUM uses the [T5 text-to-text framework](#), which is more powerful than [BERT](#). MUM not only understands text, it can also understand images. It's trained across 75 different languages and multilingual, allowing it to develop a more comprehensive understanding of and world knowledge than previous models. And MUM understands information across text and images, and can expand to more modalities like video and audio.

Take the question about hiking Mt. Fuji: MUM could be comparing two mountains, so elevation and trail information. It could also understand that, in the context of hiking, you might include things like fitness training as well as finding a guide.

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

Start now

Our platform can be plugged into any

Large language models

Our models have been trained on a wide range of data, allowing them to understand and generate text.

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://cohere.ai/>

<https://ask-rbg.ai/>

The view from 2023



ChatGPT: Optimizing Language Models for Dialogue

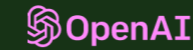
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

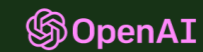
ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.



ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.



ChatGPT: Optimizing Language Models for Dialogue

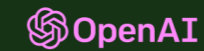
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.



ChatGPT: Optimizing Language Models for Dialogue

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

e
er
nge

uctGPT,
rompt

The view from 2023

- Have tech cos cooled down the AI hype? (Of course not)
- Have people at large become better at critically analyzing claims of “understanding language”?
- For more updates: Please join us for Stochastic Parrots Day! 17 March 2023

- 4pm-8pm CET

- <https://bit.ly/ParrotsDay23>



Stochastic Parrots Day

**March 17 8am-12pm PST/
3pm-7pm GMT**

Outline

- Problematization of the rush for scale and the “foundation models” conceptualization
- Risks associated with large (and ever larger) LLMs
- Evaluation & its prerequisites
- Value sensitive design and techniques for mitigating risk
- Dangers and responsibilities that come with working on a ‘hot topic’

Back to reality...

- Evaluation is central to
 - NLP as science
 - NLP as tech development
- Before building anything, plan out evaluation: accuracy, effectiveness, safety
- But all of the above requires knowing:
 - What it's for
 - Who will be affected

Short history of evaluation of LMs

[some citations needed]

- When LMs were used for smoothing out the output of acoustic models and translation models
 - Intrinsic evaluation: Perplexity
 - Extrinsic evaluation: WER, BLEU
- Then Bolukbasi et al 2016, Caliskan et al 2017: Evaluating word embeddings for bias
 - Then the metric becomes a goal ... see Gonen & Goldberg 2019
- Word embeddings and then LMs evaluated with benchmarks like GLUE (e.g. Devlin et al 2019) — see Raji et al 2021

So, what are LLMs for?

- I've seen a lot of bad ideas:

- LLMs as search engines

- LLMs as robo-lawyers

- LLMs as psychotherapists

- LLMs as diagnostic machines

- LLMs as stand-ins for human subjects in political science surveys

<https://bit.ly/MAIHT3k>

So what are LLMs for?

- What are some good ideas?
- Use cases where:
 - What matters is language form (content is unimportant or otherwise handled)
 - Ersatz coherence won't be misleading
 - Problematic biases/hateful content can be identified and filtered

Outline

- Problematization of the rush for scale and the “foundation models” conceptualization
- Risks associated with large (and ever larger) LLMs
- Evaluation & its prerequisites
- Value sensitive design and techniques for mitigating risk
- Dangers and responsibilities that come with working on a ‘hot topic’

From value sensitive design: Stakeholder analysis

- Direct stakeholders: people who use the technology or are involved in its development
- Indirect stakeholders: people who are affected by others' use or development of that technology
- Who are the direct stakeholders for large language models?
- Who are the indirect stakeholders?

From value sensitive design: Stakeholder analysis

- Direct stakeholders: people who use the technology or are involved in its development
- Indirect stakeholders: people who are affected by others' use or development of that technology
 - Think of a specific use case for large language models?
- Who are the direct stakeholders for that technology?
- Who are the indirect stakeholders?

Envisioning cards (Friedman et al 2011)

- Sticking with the use cases you thought of for the previous slide
- Discuss with your partner the prompt on your envisioning card
- How does that inform how you might evaluate technology for your use case?
- How does that inform how you might evaluate LLMs?

Dataset documentation: Data Statements v2

- Risk mitigation often requires knowledge of what's in a dataset
- Is this a good match for my use case?
- How might it go wrong?
- 2017: Multiple sites start creating documentation tool kits
- For language data: Data statements (Bender & Friedman 2018)
- v2 + best practices guide: <http://techpolicylab.uw.edu/data-statements/> (Bender et al 2021, McMillan-Major et al forthcoming)

Dataset documentation: Data Statements v2

SCHEMA ELEMENTS VERSION 2

- 1 HEADER
- 2 EXECUTIVE SUMMARY
- 3 CURATION RATIONALE
- 4 DOCUMENTATION FOR SOURCE DATASETS
- 5 LANGUAGE VARIETIES
- 6 SPEAKER DEMOGRAPHIC
- 7 ANNOTATOR DEMOGRAPHIC
- 8 SPEECH SITUATION AND TEXT CHARACTERISTICS
- 9 PREPROCESSING AND DATA FORMATTING
- 10 CAPTURE QUALITY
- 11 LIMITATIONS
- 12 METADATA
- 13 DISCLOSURE AND ETHICAL REVIEW
- 14 OTHER
- 15 GLOSSARY

Outline

- Problematization of the rush for scale and the “foundation models” conceptualization
- Risks associated with large (and ever larger) LLMs
- Evaluation & its prerequisites
- Value sensitive design and techniques for mitigating risk
- Dangers and responsibilities that come with working on a ‘hot topic’

Dangers and opportunities

- We're not off in a ~~corner~~ ivory tower working on theoretical things
- How we talk about what we do matters:
 - In publications and blog posts
 - To our university PR services
 - To the media
- Opt in to this, if you have time and energy. We need more experts willing to push back against the hype.

Dangers and opportunities

- Corporate entities may well pick up what we develop
 - Write thoughtful “ethical considerations” sections
- Seek chances to inform policy makers
 - And point to the ethical considerations

Thank you!



- Problematization of the rush for scale and the “foundation models” conceptualization
- Risks associated with large (and ever larger) LLMs
- Evaluation & its prerequisites
- Value sensitive design and techniques for mitigating risk
- Dangers and responsibilities that come with working on a ‘hot topic’

References

- Bender, E. M., Friedman, B., and McMillan-Major, A. (2021a). A guide for writing data statements for natural language processing. Available at <http://techpolicylab.uw.edu/data-statements/>.
- Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S., and et al (2021b). On the dangers of stochastic parrots: Can language models be too big? *Pairing 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>.
- Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.
- Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- 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.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- 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*.
- Friedman, B. and Hendry, D. (2012). The envisioning cards: A toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, page 11451148, New York, NY, USA. Association for Computing Machinery.
- Friedman, B., Nathan, L. P., Kane, S., and Lin, J. (2011). Envisioning cards. University of Washington, Seattle, WA, USA. Available at: envisioningcards.com.
- Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hanna, A. and M., T. (2020). Against scale: Provocations and resistances to scale thinking. In *CSCW Workshop 20*, Virtual.
- 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.
- Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. (2021). Ai and the everything in the whole wide world benchmark.
- 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>