Large Language Modeling: From Stochastic Parrots To Today

Angie McMillan-Major, UW Language Learning Center

For LING574
May 28, 2025

About myself

- UW Linguistics
  - Undergrad
    - Language documentation, field linguistics, and language reclamation
  - CLMS
  - PhD
    - Hugging Face 🤗 Internship
- UW Language Learning Center

Documentation and Transparency

- **How can we make NLP datasets and models more transparent and accessible?**
    - Data Statements for NLP: Documentation schema for language data
- **What are the dangers that motivate the need for transparency and accessibility?**
    - On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜
    - Emily M. Bender, Timnit Gebru, Angie McMillan-Major, and Margaret Mitchell
        - Risks and mitigation strategies

Questions for you

What risks you think are most relevant today?

Are there risks that you think haven’t gone as we envisioned them at the time?

Are there new risks we need to consider as the field has evolved in the last half decade?

Other thoughts?

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

Stochastic Parrots considerations

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 NLP 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?

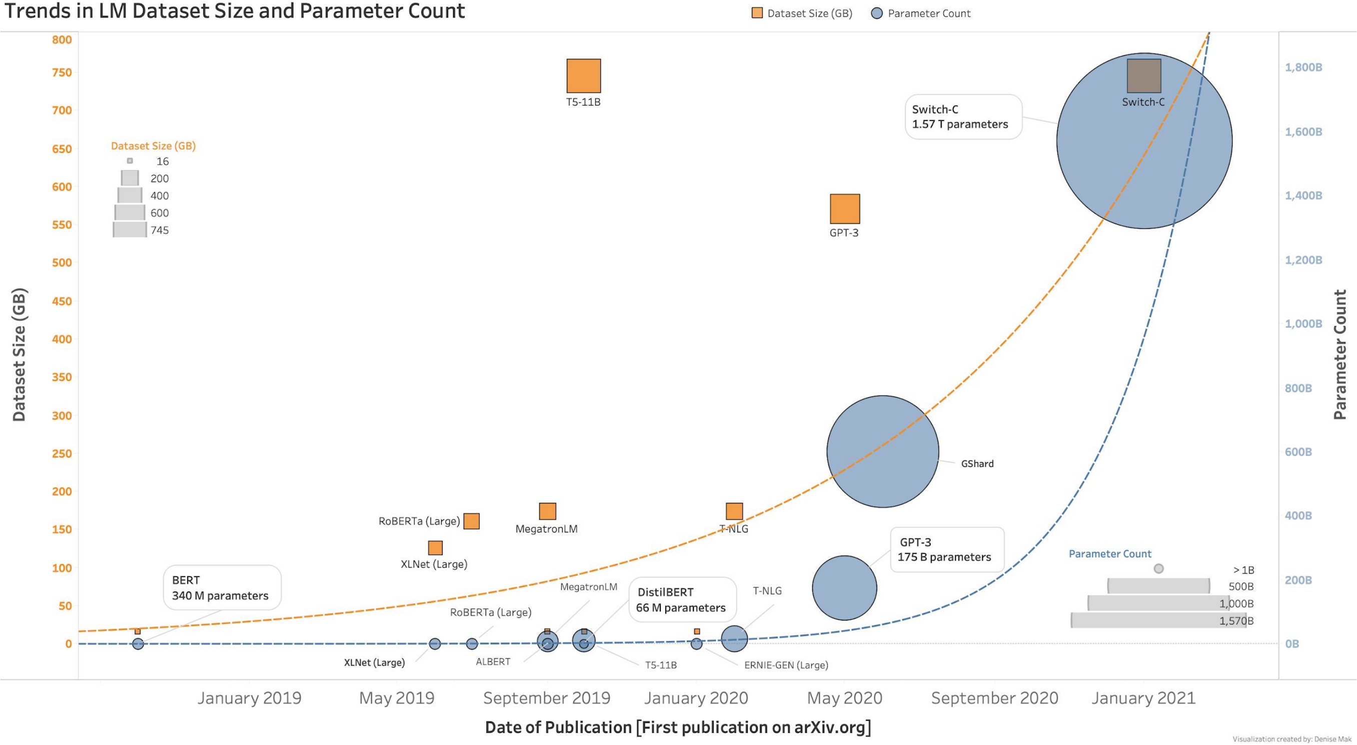
Overview

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

Brief History of LMs

- LM: A system trained to do string prediction
  - *What word comes* ___? *What word* [MASK] *here*?
- Proposed by Shannon in 1949, based on Markov's research from the early 1900's, but implemented for ASR, MT, etc. in early 80's
  - N-grams and various neural architectures until Transformers
- Big takeaways
  - Pattern of achieving 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

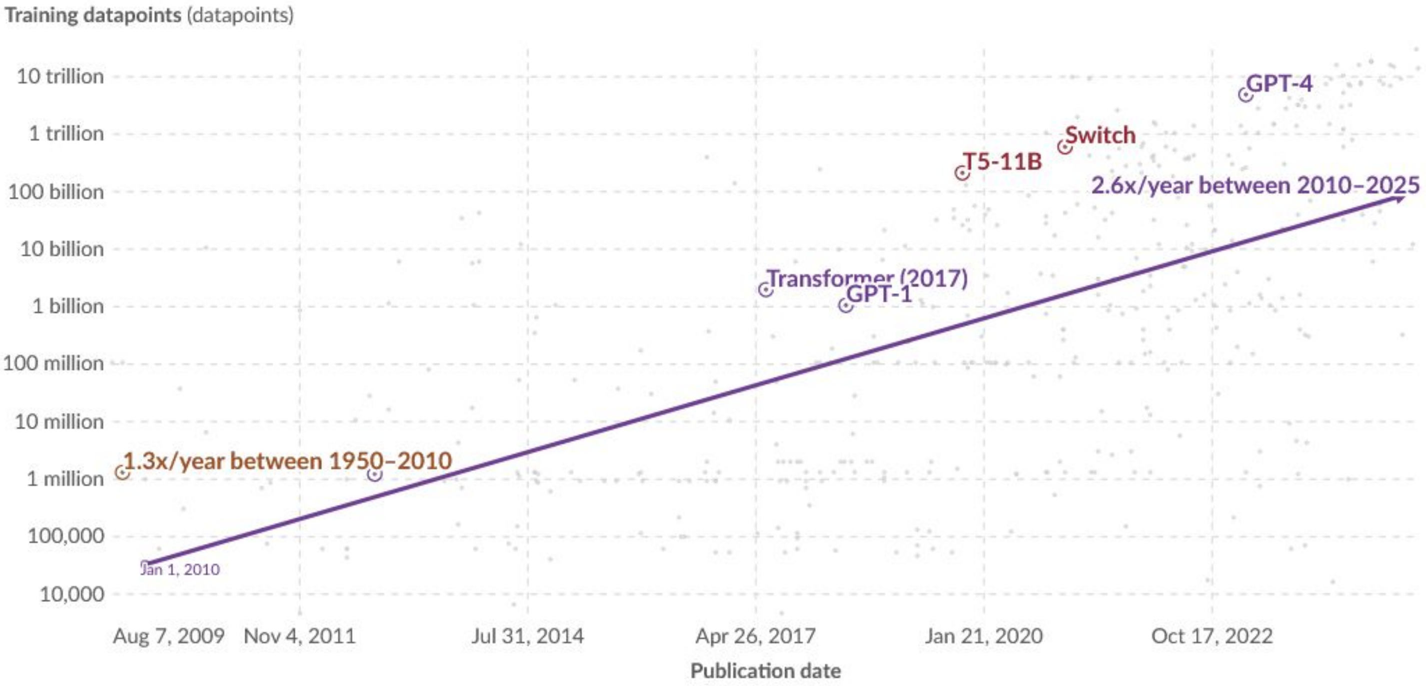
How big is big?



Visualization created by: Denise Mak

Special thank you to Denise Mak for graph design

Growth Rate Acceleration



Epoch (2025) – with major processing by Our World in Data

What are the risks?

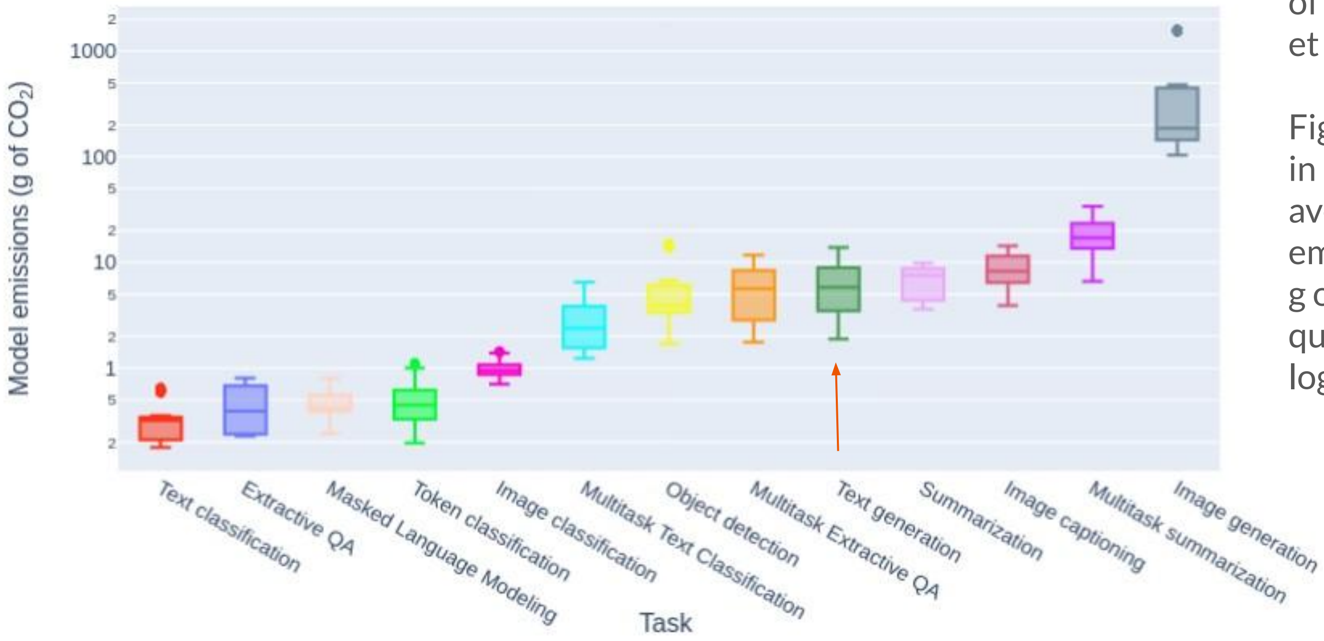
Environmental Costs
Financial Inaccessibility

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 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

Energy Costs Beyond Training



Power Hungry Processing: ⚡Watts⚡Driving the Cost of AI Deployment? (Luccioni et al, 2024)

Figure 1: The tasks examined in [their] study and the average quantity of carbon emissions they produced (in g of $CO2eq$) for 1,000 queries. N.B. The y axis is in logarithmic scale.

Mitigation Efforts

- Renewable energy sources
  - But still incur a cost on the environment
  - Still 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)
  - ACL Rolling Review Checklist recommends reporting Total Computation Budget (GPU hours)

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

Interrogating Infrastructure

Materiality (Borning, Friedman, and Gruen, 2018; Borning, Friedman, and Logler, 2020)

UW Patas Cluster - Tower Data Center

- Was ENERGY STAR certified 2013-2023
- 520-ton cooling tower, which utilizes ambient outside air, a fan and a pump to reject waste heat from the data center
- Automation to reduce motorized usage of cooling systems

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 subjected 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

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
- Efforts towards post-hoc documentation:
  - Bandy and Vincent (2021) - retrospective datasheet for BookCorpus
  - Dodge et al (2021) - examining the contents of C4

What are the risks?

Research Trajectories

Research Time is a 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)
  - Raji et al (2021) construct validity issues of benchmarks for “general purpose”
- LMs trained only on linguistic form don’t have access to meaning (Bender & Koller 2020)
- Are we actually learning about machine language understanding?

Towards More Thorough Research Practices

- On the Gap between Adoption and Understanding in NLP (Bianchi and Hovy, 2021)
  - Current research practices disincentivize publishing negative (or just not state of the art) research results

- On "Scientific Debt" in NLP: A Case for More Rigour in Language Model Pre-Training Research (Nityasya et al., 2023)
  - Current research practices manipulate too many variables in each experiment to isolate and understand the variables

What are the risks?

Abusive Language and Synthetic Data

Stochastic 🦜

- Human-human interaction is co-constructed and leads to a shared model of the world (Reddy 1979, Clark 1996)

- An LM is 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 Identified in 2021

- 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)

Human Labor Costs of AI

The Exploited Labor Behind Artificial Intelligence (Williams, Miceli, and Gebru, 2022), especially the labor and trauma of human content moderation, annotation, and filtering:

- The Trauma Floor: The secret lives of Facebook moderators in America | The Verge
- OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive | TIME
- The horrors experienced by Meta moderators: ‘I didn’t know what humans are capable of’ | EL PAÍS
- The Low-Paid Humans Behind AI’s Smarts Ask Biden to Free Them From ‘Modern Day Slavery’ | WIRED

Risk Mitigation 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

Data Statements for NLP

Documentation schema for language datasets: https://techpolicylab.uw.edu/data-statements/

- Version 1: Bender and Friedman (2018)
  - Original schema
- Version 2: guide (2021) and McMillan-Major, Bender, and Friedman (2024)
  - Dataset documentation refined by scientific community engagement
- Version 3: guide (2024) and McMillan-Major (2023)
  - Dataset documentation and creation with best practices for language community dataset development

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?

Policy Around the World

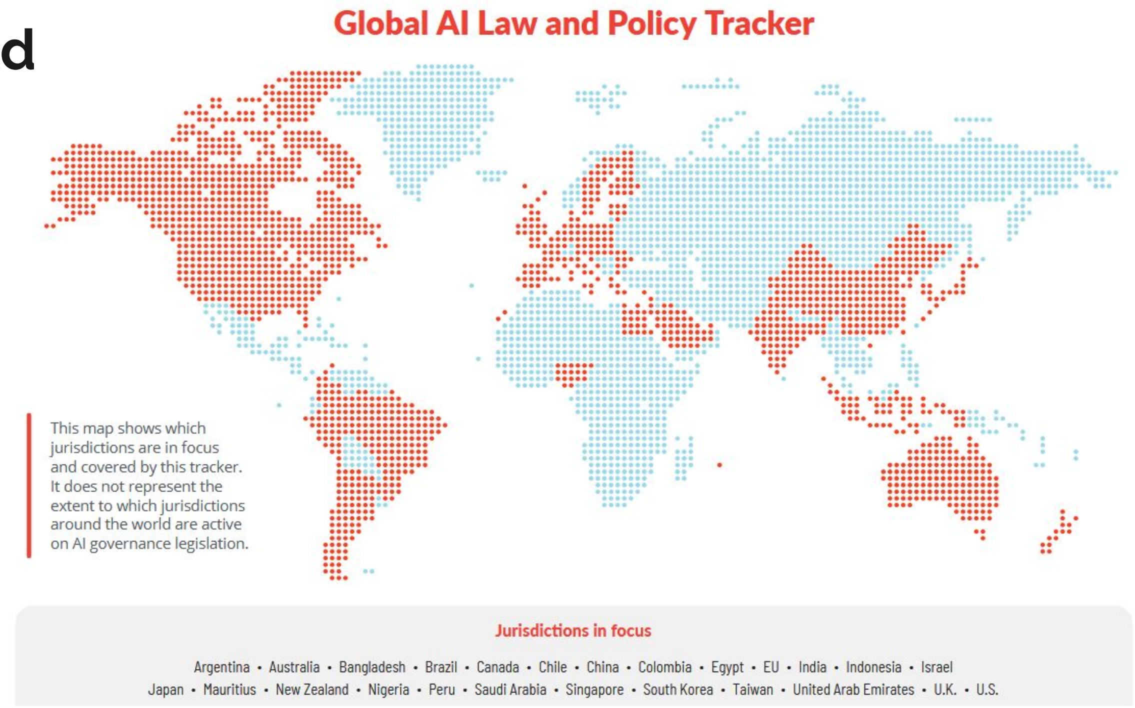
GDPR (2018)

- Rights of the data subject to their personal data

EU AI Act (2024)

- Tasks classified based on risk, with highest risk tasks subject to regulation

Global AI Law and Policy Tracker from IAPP: https://iapp.org/media/pdf/resource_center/global_ai_law_policy_tracker.pdf



Global AI Law and Policy Tracker

This map shows which jurisdictions are in focus and covered by this tracker. It does not represent the extent to which jurisdictions around the world are active on AI governance legislation.

Jurisdictions in focus

Argentina • Australia • Bangladesh • Brazil • Canada • Chile • China • Colombia • Egypt • EU • India • Indonesia • Israel
Japan • Mauritius • New Zealand • Nigeria • Peru • Saudi Arabia • Singapore • South Korea • Taiwan • United Arab Emirates • U.K. • U.S.

Questions we ended with in 2021

Are ever larger language models inevitable or necessary?

What costs are associated with this research direction and what should we consider before pursuing it?

Do the field of NLP 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?

Questions for you

What risks you think are most relevant today?

Are there risks that you think didn't go as we envisioned them at the time?

Are there new risks we need to consider as the field has evolved in the last half decade?

Other thoughts?

Thank you!

References

Please see the Stochastic Parrots paper for the full bibliography.

Claude Elwood Shannon. 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana.

“Data Page: Exponential growth of datapoints used to train notable AI systems”, part of the following publication: Charlie Giattino, Edouard Mathieu, Veronika Samborska, and Max Roser (2023) - “Artificial Intelligence”. Data adapted from Epoch. Retrieved from https://ourworldindata.org/grapher/exponential-growth-of-datapoints-used-to-train-notable-ai-systems [online resource]

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 3645--3650.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. Commun. ACM 63, 12 (Nov. 2020), 54--63. https://doi.org/10.1145/3381831

Kadan Lottick, Silvia Susai, Sorelle A. Friedler, and Jonathan P. Wilson. 2019. Energy Usage Reports: Environmental awareness as part of algorithmic accountability. arXiv:1911.08354 [cs.LG]

References

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. Journal of Machine Learning Research 21, 248 (2020), 1--43. http://jmlr.org/papers/v21/20-312.html

Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. Power Hungry Processing: ⚡Watts⚡ Driving the Cost of AI Deployment? In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery, New York, NY, USA, 85–99. https://doi.org/10.1145/3630106.3658542

Borning, A., Friedman, B., and Gruen, D. What pushes back from considering materiality in it? In *Proceedings of the 2018 Workshop on Computing Within Limits.* ACM, New York, NY.

Borning, A., Friedman, B. and Logler, N., 2020. The 'invisible' materiality of information technology. *Communications of the ACM*, *63*(6), pp.57-64.

David Anthoff, Robert J Nicholls, and Richard SJ Tol. 2010. The economic impact of substantial sea-level rise. Mitigation and Adaptation Strategies for Global Change 15, 4 (2010), 321--335.

References

Marlon Twyman, Brian C Keegan, and Aaron Shaw. 2017. Black Lives Matter in Wikipedia: Collective memory and collaboration around online social movements. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 1400--1412.

Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 4658--4664. https://doi.org/10.18653/v1/P19-1459

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial Filters of Dataset Biases. In Proceedings of the 37th International Conference on Machine Learning.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).* https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 5185--5198. https://doi.org/10.18653/v1/2020.acl-main.463

Federico Bianchi and Dirk Hovy. 2021. On the Gap between Adoption and Understanding in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3895–3901, Online. Association for Computational Linguistics.

Made Nindyatama Nityasya, Haryo Wibowo, Alham Fikri Aji, Genta Winata, Radityo Eko Prasojo, Phil Blunsom, and Adhiguna Kuncoro. 2023. On “Scientific Debt” in NLP: A Case for More Rigour in Language Model Pre-Training Research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8554–8572, Toronto, Canada. Association for Computational Linguistics.

Reddy, M., 1979. The conduit metaphor. *Metaphor and thought*, *2*, pp.285-324.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

References

Kris McGuffie and Alex Newhouse. 2020. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. Technical Report. Center on Terrorism, Extremism, and Counterterrorism, Middlebury Institute of International Studies at Monterrey. https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting Training Data from Large Language Models. arXiv:2012.07805 [cs.CR]

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Adrienne Williams, Milagros Miceli and Timnit Gebru. 2022. The Exploited Labor Behind Artificial Intelligence. *Noēma*. https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/

Billy Perrigo. 2023, January 18. Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Casey Newton. 2019, February 25. The Trauma Floor: The secret lives of Facebook moderators in America. *The Verge*. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

References

Josep Catà Figuls. 2024, January 28. The horrors experienced by Meta moderators: ‘I didn’t know what humans are capable of’. *El País.* https://english.elpais.com/economy-and-business/2024-01-29/the-horrors-experienced-by-meta-moderators-i-didnt-know-what-humans-are-capable-of.html

Caroline Haskins. 2024, May 22. The Low-Paid Humans Behind AI’s Smarts Ask Biden to Free Them From ‘Modern Day Slavery’. *WIRED*. https://www.wired.com/story/low-paid-humans-ai-biden-modern-day-slavery/

Abeba Birhane and Vinay Uday Prabhu. 2021. Large Image Datasets: A Pyrrhic Win for Computer Vision?. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1537--1547

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604. https://aclanthology.org/Q18-1041/

Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2024. Data Statements: From Technical Concept to Community Practice. *ACM J. Responsib. Comput*. 1, 1, Article 1 (March 2024), 17 pages. https://doi.org/10.1145/3594737