Ethics in NLP

Instructor: Lara J. Martin (she/they)

TA: Omkar Kulkarni (he)

https://laramartin.net/NLP-class/

Learning Objectives

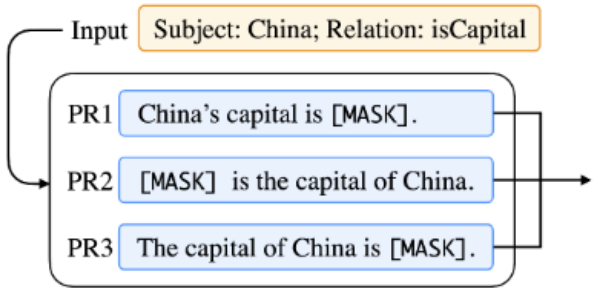
Identify ethical issues of LLMs/transformers from various lenses (social, environmental, legal, economic, etc.) by…

- Extracting them from the Stochastic Parrots paper
- Extending them with your own perspectives

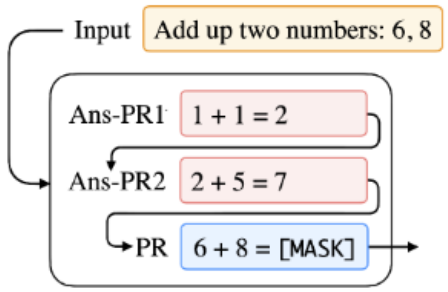
Determine how these issues apply to any LM

Review: Multi-prompt “Learning”

Multiple unanswered prompts



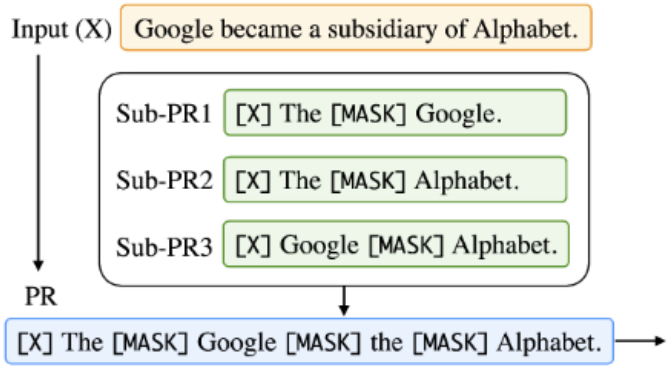
(a) Prompt Ensembling.



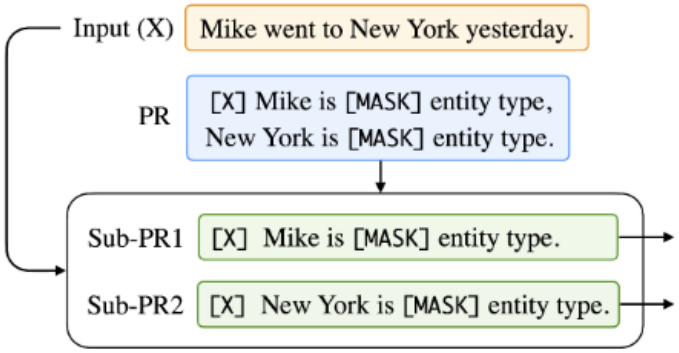
(b) Prompt Augmentation.

Demonstration learning

Merging subprompts



(c) Prompt Composition.



(d) Prompt Decomposition.

For multi-label predictions

“ ” for input text, “ ” for prompt, “ ” for answered prompt. “ ” for sub-prompt.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Comput. Surv. 55, 9, Article 195 (September 2023), 35 pages. https://doi.org/10.1145/3560815

Review: Self-Criticism

LLM “ruminates” on its output to try to come up with better output

(Along with chain-of-thought) Precursor to reasoning models that are finetuned to do this automatically

```
Question: Who was the third president of the United States?
Here are some brainstormed ideas: James Monroe
Thomas Jefferson
John Adams
Thomas Jefferson
George Washington
Possible Answer: James Monroe
Is the possible answer:
 (A) True
 (B) False
The possible answer is:
```


Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, *et al.* 2022. Language Models (Mostly) Know What They Know. arXiv: 2207.05221

Review: Meta-Prompting

Prompting to generate prompts

Meta-prompts…

- “encode what the LLM should “do” given different descriptions of the same task” [1]
- “will return the relevant outputs for an arbitrary task, provided that the task description is provided as an input” [1]

[1] Adrian de Wynter, Xun Wang, Qilong Gu, Si-Qing Chen. 2023. On Meta-Prompting. arXiv: 2312.06562

Stochastic Parrots

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623. https://doi.org/10.1145/3442188.3445922

Ethical Issues

1. Environmental
2. Financial
3. Diversity
4. Static Data
5. Bias
6. Accountability
7. Lack of Understanding
8. Subjective Coherence
9. Harms

+ 10. Mitigation Strategies

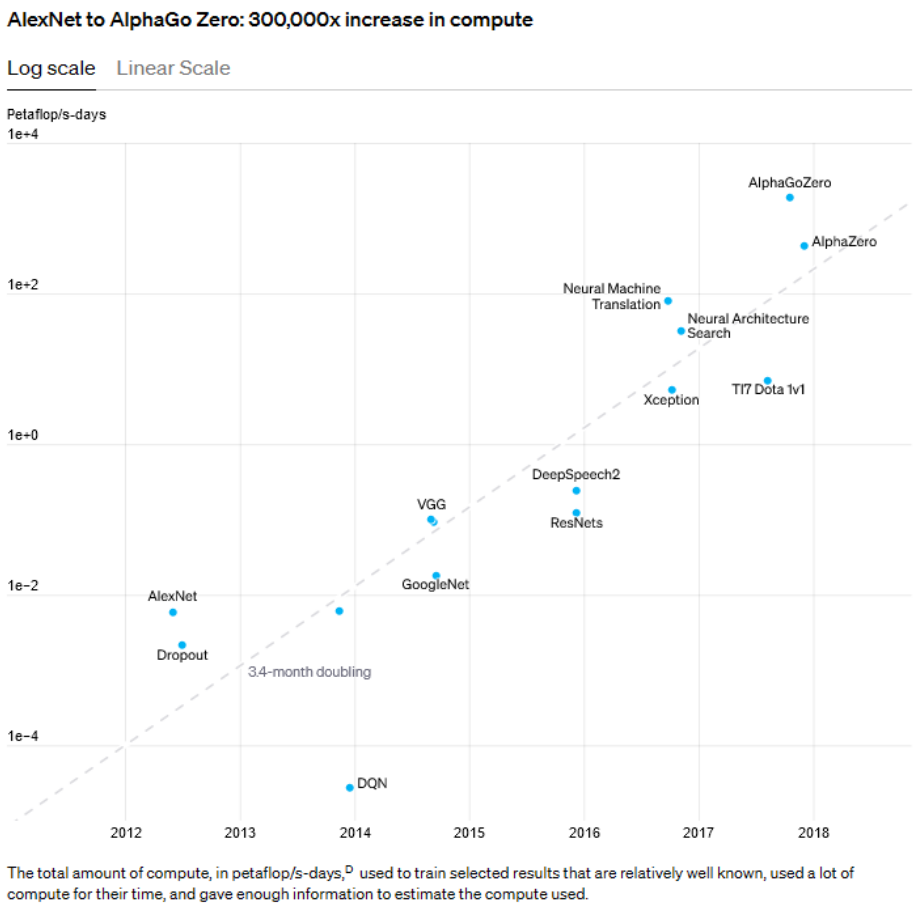
1) Environmental

1. Training a model produces CO2 gas
2. Data centers take all the renewable energy
3. Marginalized areas are most impacted but least likely to benefit from the tools

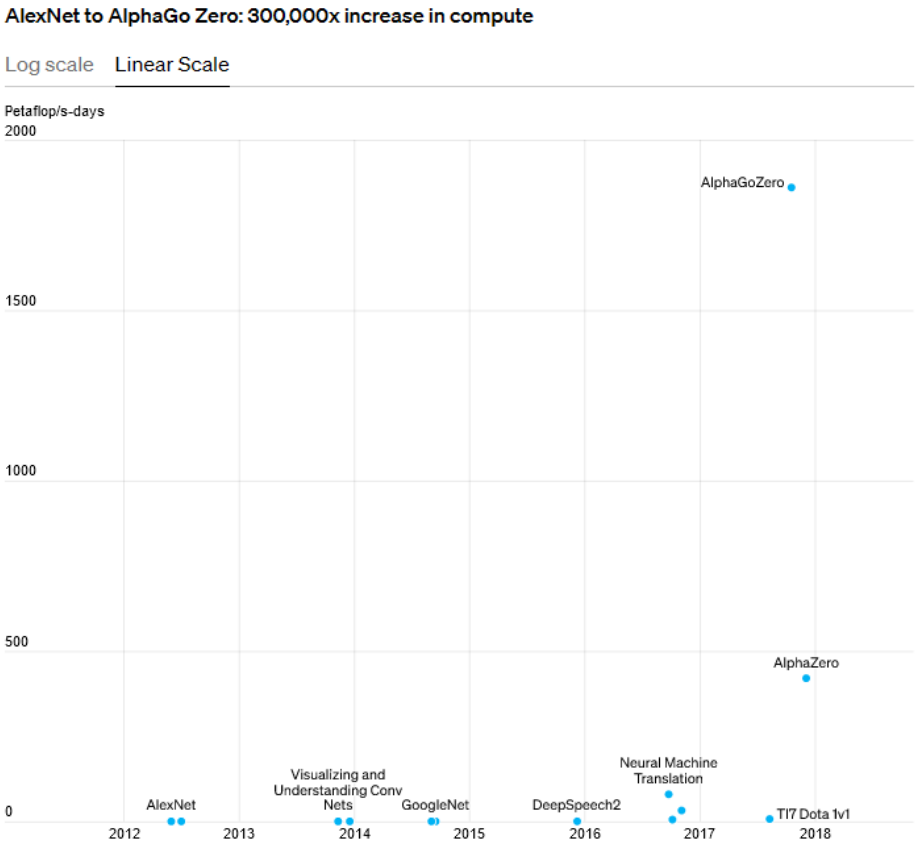
Mitigation:

1. Low-energy models? Hardware that specializes
2. Treating efficiency as a metric (in addition to accuracy, etc.)

Energy of Models



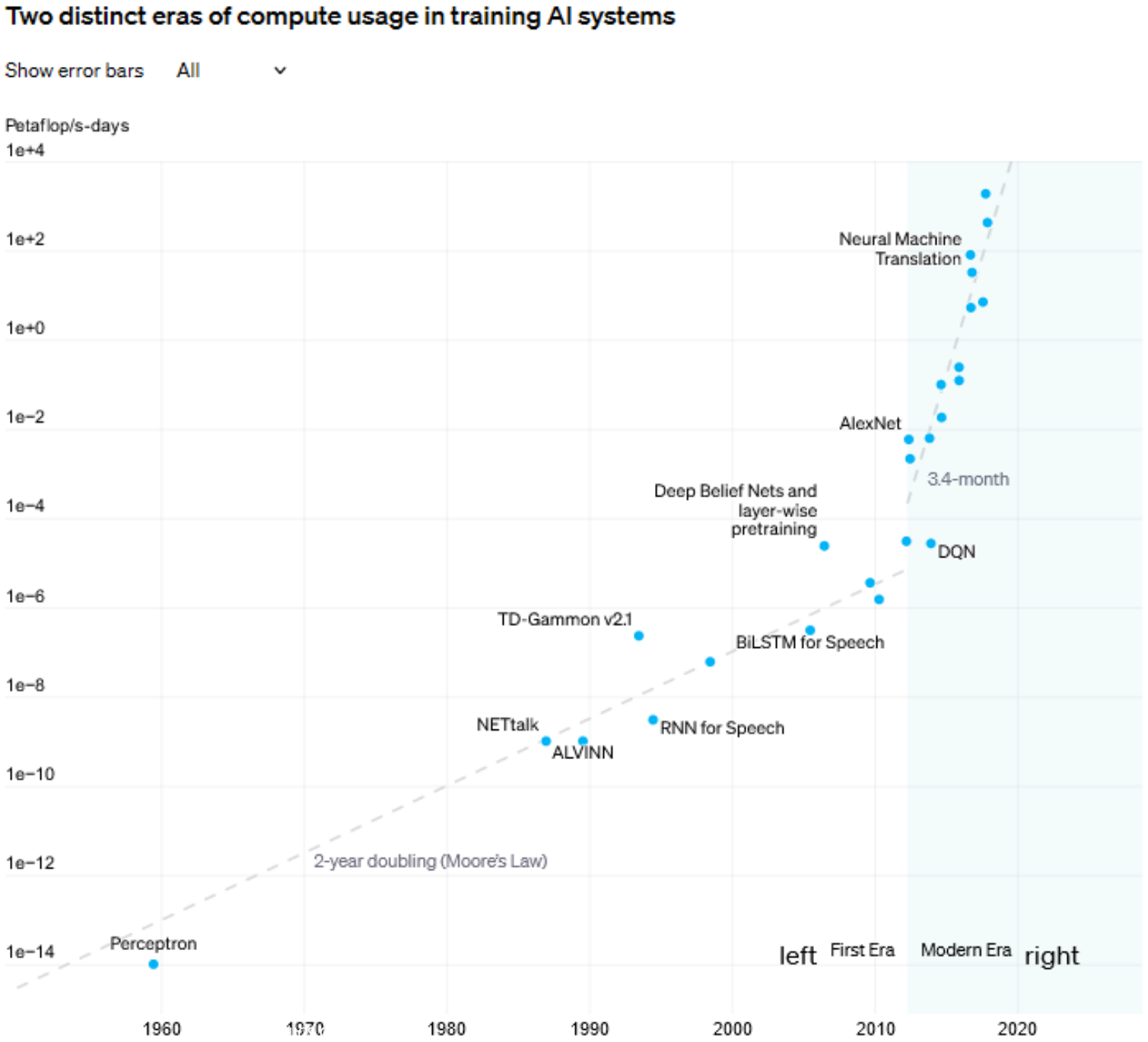
The total amount of compute, in petaflop/s-days,[D] used to train selected results that are relatively well known, used a lot of compute for their time, and gave enough information to estimate the compute used.



The total amount of compute, in petaflop/s-days,[D] used to train selected results that are relatively well known, used a lot of compute for their time, and gave enough information to estimate the compute used.

https://openai.com/research/ai-and-compute

Energy Shift



https://openai.com/research/ai-and-compute

2) Financial

1. Correlation between BLEU and cost to get to that score
2. Training vs inference – if a model is used frequently, inference cost matters
3. Barriers of entry for research
4. Wealthiest institutions benefit from models the most

Mitigations:

1. Share cloud computing resources (GPU clusters & APIs)
2. Open source models or institutionally-shared models

3) Diversity

1. Much of the data comes from young males on the internet
2. Older people are more likely to use blogs but these are less connected (less SEO)
3. More data doesn't mean more diversity

Mitigations:

1. Be more thorough in dataset prep; diverse collection of data from different groups
2. Reduce size if it can't be prepared properly
3. Spend more time collecting data for the specific task

Near Duplicates in Data

Dataset	Example	Near-Duplicate Example
Wiki-40B	\n_START_ARTICLE_\nHum Award for Most Impactful Character \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]	\n_START_ARTICLE_\nHum Award for Best Actor in a Negative Role \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]
LM1B	I left for California in 1979 and tracked Cleveland 's changes on trips back to visit my sisters .	I left for California in 1979 , and tracked Cleveland 's changes on trips back to visit my sisters .
RealNews	KUALA LUMPUR (Reuters) - Roads in Southeast Asia have been getting a little louder lately as motorcycle makers, an aspiring middle class and easy bank credit come together to breed a new genus of motorcyclists – the big-bike rider. [...]	A visitor looks at a Triumph motorcycle on display at the Indonesian International Motor Show in Jakarta September 19, 2014. REUTERS/Darren Whiteside\n KUALA LUMPUR (Reuters) - Roads in Southeast Asia have been getting a little [...] big-bike rider. [...]
C4	Affordable and convenient holiday flights take off from your departure country, "Canada". From May 2019 to October 2019, Condor flights to your dream destination will be roughly 6 a week! Book your Halifax (YHZ) - Basel (BSL) flight now, and look forward to your "Switzerland" destination!	Affordable and convenient holiday flights take off from your departure country, "USA". From April 2019 to October 2019, Condor flights to your dream destination will be roughly 7 a week! Book your Maui Kahului (OGG) - Dubrovnik (DBV) flight now, and look forward to your "Croatia" destination!

Table 1: Qualitative examples of near-duplicates identified by NEARDUP from each dataset. The similarlity between documents is highlighted. Note the small interspersed differences that make exact duplicate matching less effective. Examples ending with "[...]" have been truncated for brevity.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022). Deduplicating Training Data Makes Language Models Better. *Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, 8424–8445. https://doi.org/10.18653/v1/2022.acl-long.577

4) Static Data

1. LLMs are trained at a certain point in time → value locked
2. Expensive to retrain/finetune
3. Current events used to update the models tend to be more violent (things that the news covers)

Mitigations:

1. Models can retrieve more recent information (RAG)
2. Users can rate factuality / add more recent information

5) Bias

1. Because models are statistical, they are the voice of “everyone” – simultaneously sound like no one person but also the majority group
2. Trained on web data = pick up biases of humans
3. Toxicity metrics don’t take into consideration the speaker

Mitigations:

1. Compiling data that represents different perspectives (diverse data)
2. Research ways to reduce harm
3. Use multiple models and find consensus instead of relying on a single model’s bias

Reporting Bias

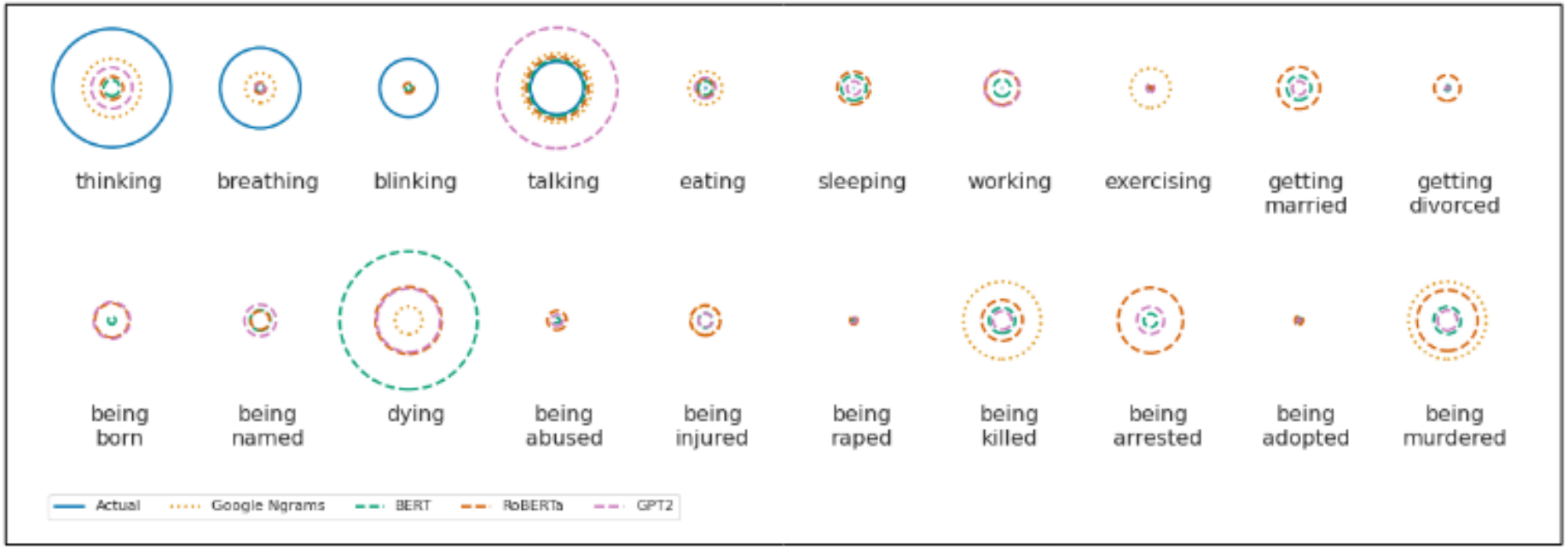


Figure 1: Frequency of actions performed or occurring to people during their lifetime from very frequent (daily), through once in a lifetime events, to very rare (don't happen to most people). Note that actual frequencies of rare events are too small to show. See Appendix A for the exact frequencies.

Shwartz, V., & Choi, Y. (2020). Do Neural Language Models Overcome Reporting Bias? *International Conference on Computational Linguistics (COLING)*, 6863–6870. https://doi.org/10.18653/v1/2020.coling-main.605

6) Accountability

1. If a model outputs something harmful, the model can't be held accountable legally
2. Developers can't know everything the model was trained on so they're not accountable
3. How much is the user's prompt influencing the output of the model?

Mitigations:

1. Having guardrails to avoid the situations in the first place
2. Warning users on how to use the model
3. License for training/releasing models

7) Lack of Understanding

1. LLMs just copy humans but don’t truly understand
2. Humans ascribe too much meaning to the LLM

Mitigations:

1. Teaching people how models work
2. Interpretability & Explainable AI

8) Subjective Coherence

1. Machine outputs sentence that sounds good but is inaccurate
2. Lacking cultural understanding

Mitigations:

1. Admitting uncertainty

9) Harms

1. Harmful text in training data can spread harm to other contexts
2. Propagate harmful stereotypes (even bigger issue if models are trained on harmful generated data)

Mitigations:

1. Labeling harmful content (or AI generated in general)
2. Put more effort into planning (goals, values, motivations, users, stakeholders are considered)
3. Reverse engineer hypothetical failures (pre-mortem)

Issues of all LMs

How many of these are relevant to smaller LMs?