Ethics in NLP

CMSC 473/673 - NATURAL LANGUAGE PROCESSING

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: What is a foundation model?

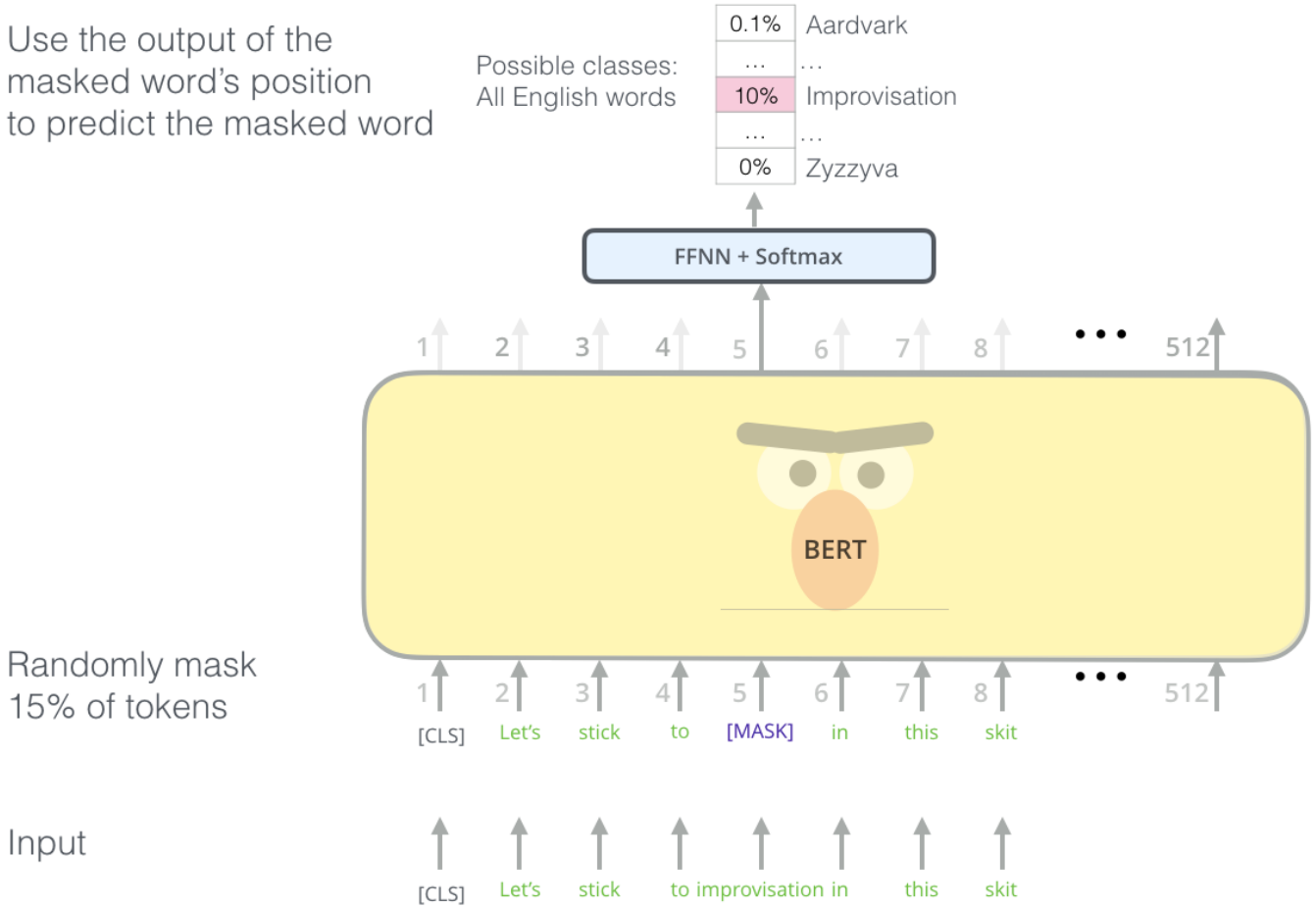
A model that captures “foundation” or core information about a modality (e.g., text, speech, images)

Pretrained on a large amount of data & able to be finetuned on a particular task

Self-supervised

All non-finetuned large language models (LLMs) are foundation models

Review: BERT (Devlin et al. 2019)



http://jalammar.github.io/illustrated-bert/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Review: GPT Family

- Decoder-only
  - Input: Text sequence
  - Goal: Predict the next word given the previous ones
- How to use:
  - Ask GPT* to continue from a prompt.
  - Finetune smaller GPTs for more customized generation tasks.
    - ChatGPT cannot be finetuned since it is already finetuned
  - Use OpenAI's API to get them to fine-tune GPT-3 for you.

Review: T5 Family of Models

- Encoder-decoder
  - Input: Text sequence with random word spans deleted
  - Goal: Generate the deleted word spans
- How to use:
  - Finetune smaller ones for either generation or classification tasks.
  - Prompt tuning (train a sequence of embedding which get prefixed to the input)

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. Needs water cooling; can deplete water sources
2. Lots of electricity – possibly powered by fossil fuels (climate change)
3. Manufacturing of hardware that can handle the models
4. Generating heat

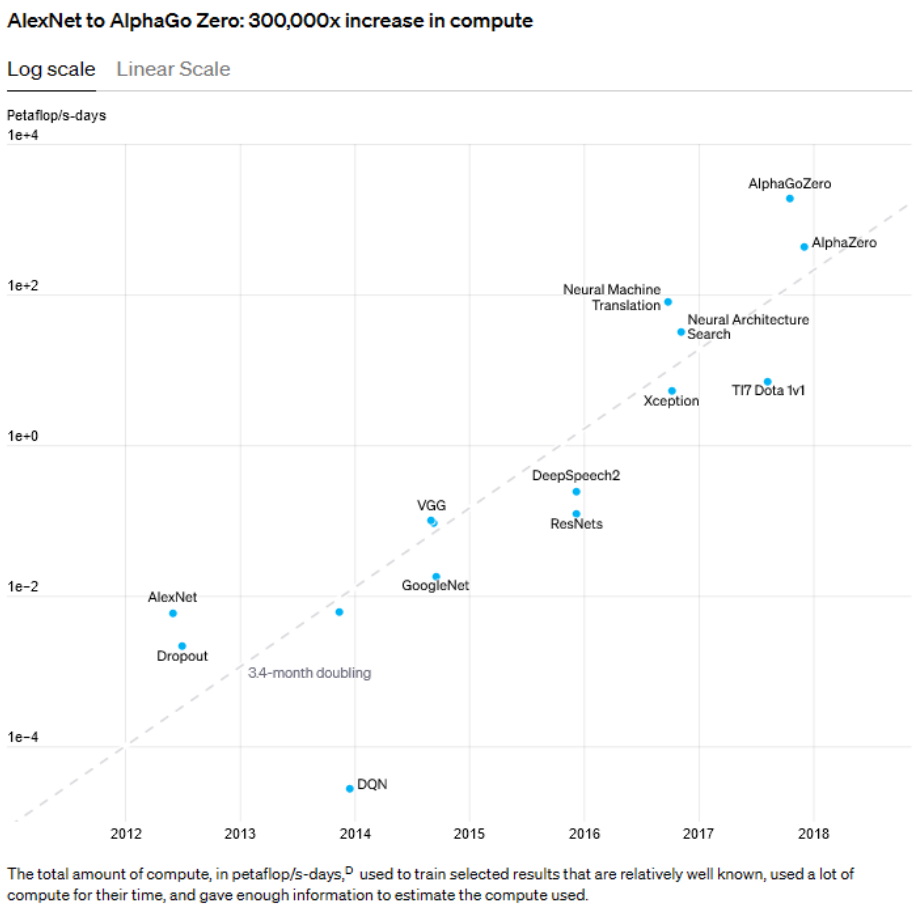
Mitigations:

Not default to using LLM when not requested/wanted

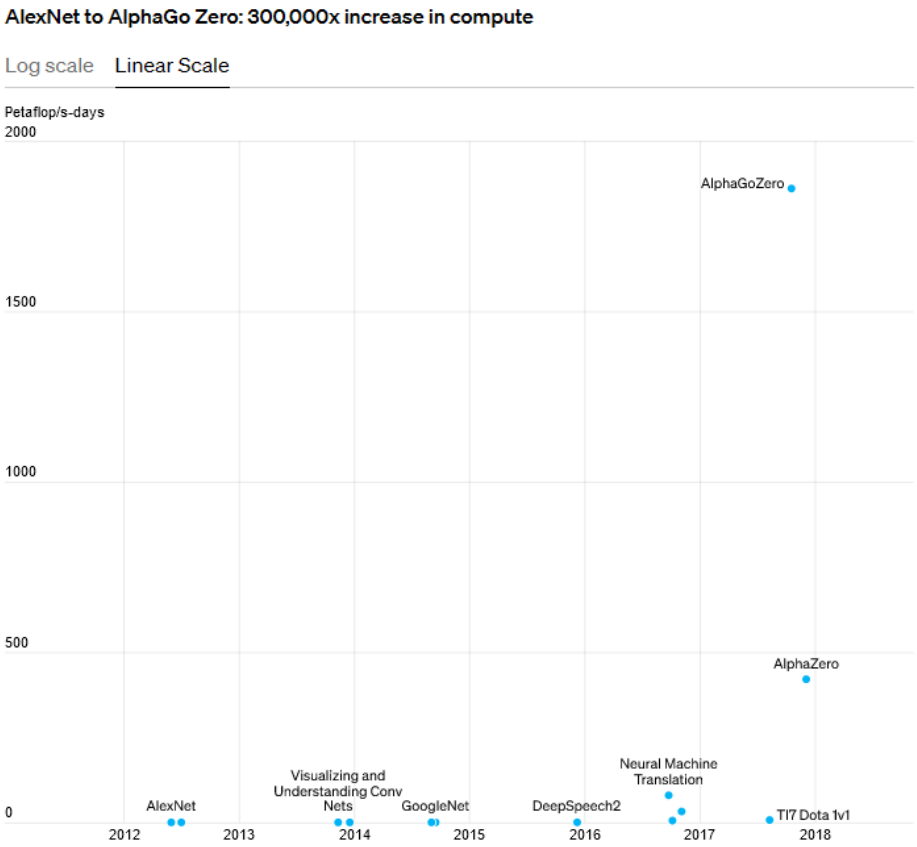
Make more energy-efficient models

Explore other ways to cool

Energy of Models



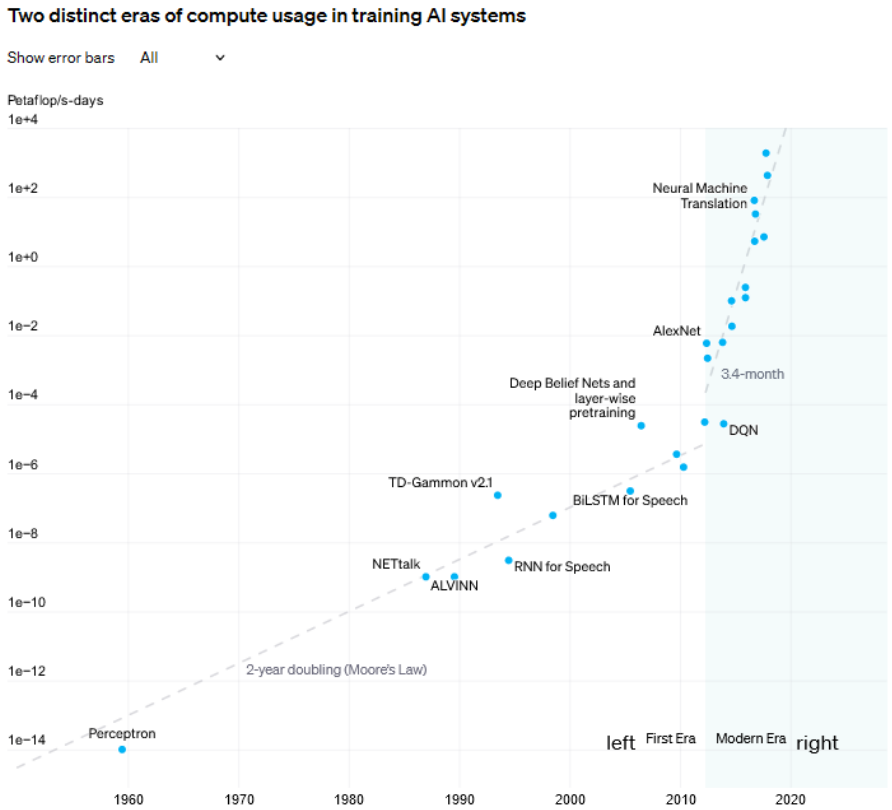
The total amount of compute, in petaflop/s-days,[D] used to train selected results that are relatively well known, used a lot of compute for their time, and gave enough information to estimate the compute used.



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. As models become bigger, they become more expensive
   A. Electricity Costs, other utilities
   B. Hardware
   C. Warehouses
2. Makes other applications more expensive
3. Economic inequality
   A. It costs to use them
   B. It used to be that you had to buy the hardware (not much better)

   Mitigations:

   Prioritize same task without LLM (Occam's Razor)

   Consider if it's worth including an LLM

3) Diversity

1. Training on the internet prioritizes white supremacy, etc.
   A. Online public spaces have minorities underrepresented
2. Filtering can be harmful because the underrepresented topics are removed entirely or because they use certain keywords
3. Internet access; who posts online

Mitigations:

Have more human input who belong to the minority groups

4) Static Data

1. When prompted, it might pull from old data but the user might not know
2. If it wasn't in the training data, it wouldn't know about it at all
3. Unable to change perspectives on people and events

Mitigations:

Mention when the data was collected and that it's outdated

Users can give feedback on updated information

5) Bias

1. Trained on social media that is Western, cis/het white men (what the data is one)
2. The companies aren't explaining the harms & even if they are, the people aren't reading it
3. Social and political decisions in collecting data (how the data is made)

Mitigations:

Choose a variety of data from various sources

Keep up to date with social changes

Reporting Bias

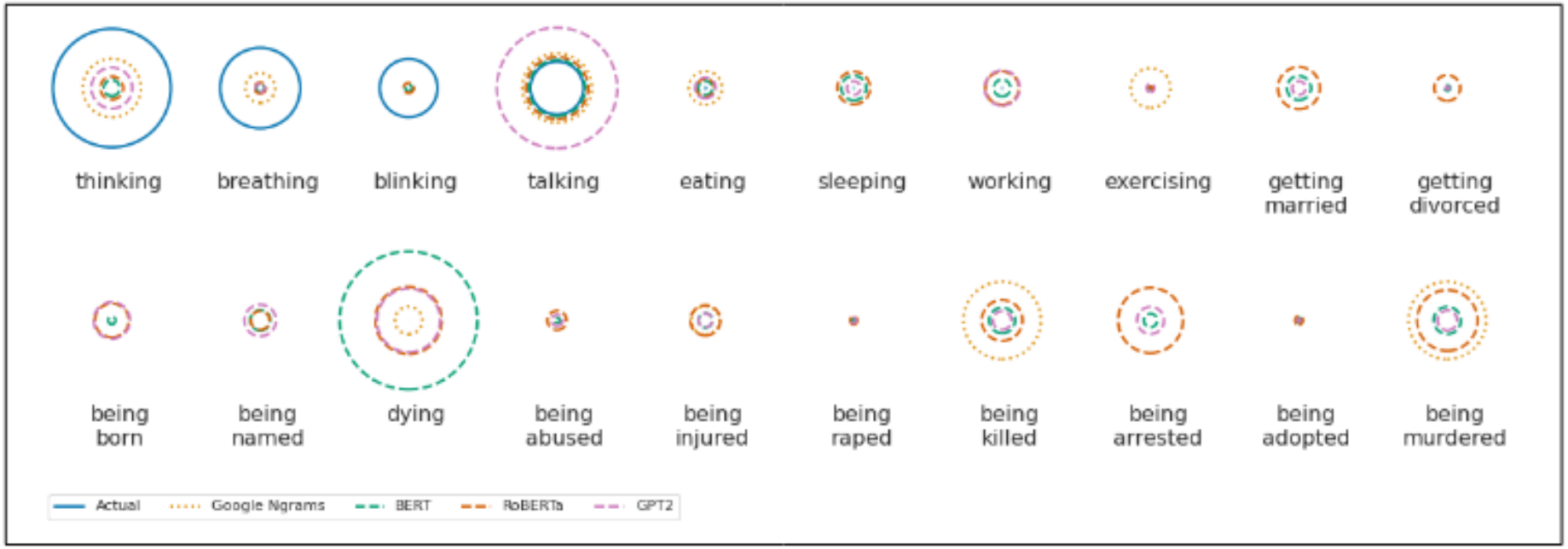


Figure 1: Frequency of actions performed or occurring to people during their lifetime from very frequent (daily), through once in a lifetime events, to very rare (don't happen to most people). Note that actual frequencies of rare events are too small to show. See Appendix A for the exact frequencies.

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

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

6) Accountability

1. Developers are accountable for the data, the model architecture
   A. But they can't know all the data → impossible to document it all
   B. It needs to be curated & documented where it's from
2. Who takes the blame if the model makes a mistake? How is that mistake explained to people?

Mitigations:

Incentives to document the data; make annotation tools to make it easier

Devs need to understand the harms and communicate issues

7) Lack of Understanding

1. As the model size increases, the LLMs struggle with understanding the training data
2. It's not actually "understanding" what the user wants → if it gets it wrong, then this can be an issue

Mitigations:

Take the generation with a grain of salt & explain that to the users

Other "reasoning" models and techniques can help → explainable AI

8) Subjective Coherence

1. Interpretation depends on the user
2. Generated text can be completely nonsensical
3. Can lose it's train of thought
   A. Cannot make sense of events → can't deal with causality
   B. Disorganized thoughts

Mitigations:

Generate shorter bits of text at a time

Ethical finetuning (guardrails)

Provide a frame for the response so it's less confident when it's irrational

Prioritize facts over arguments

Only use credible sources

9) Harms

1. Gender bias & other harms for users from different groups
2. Output can be highly skewed → misinformation
3. Overreliance
4. Misuse of LLMs (e.g., “fake news”)
5. Cultural differences not captured in translations

Mitigations:

Curated datasets to minimize harm

Evaluate & check for harms directly

Get minorities to evaluate

Include in terms of service & explain how to use the model

Explicitly mention source of information