CMSC 473/673
Natural Language Processing

Instructor: Lara J. Martin (she/they)

TA: Duong Ta (he)

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 foundational model?

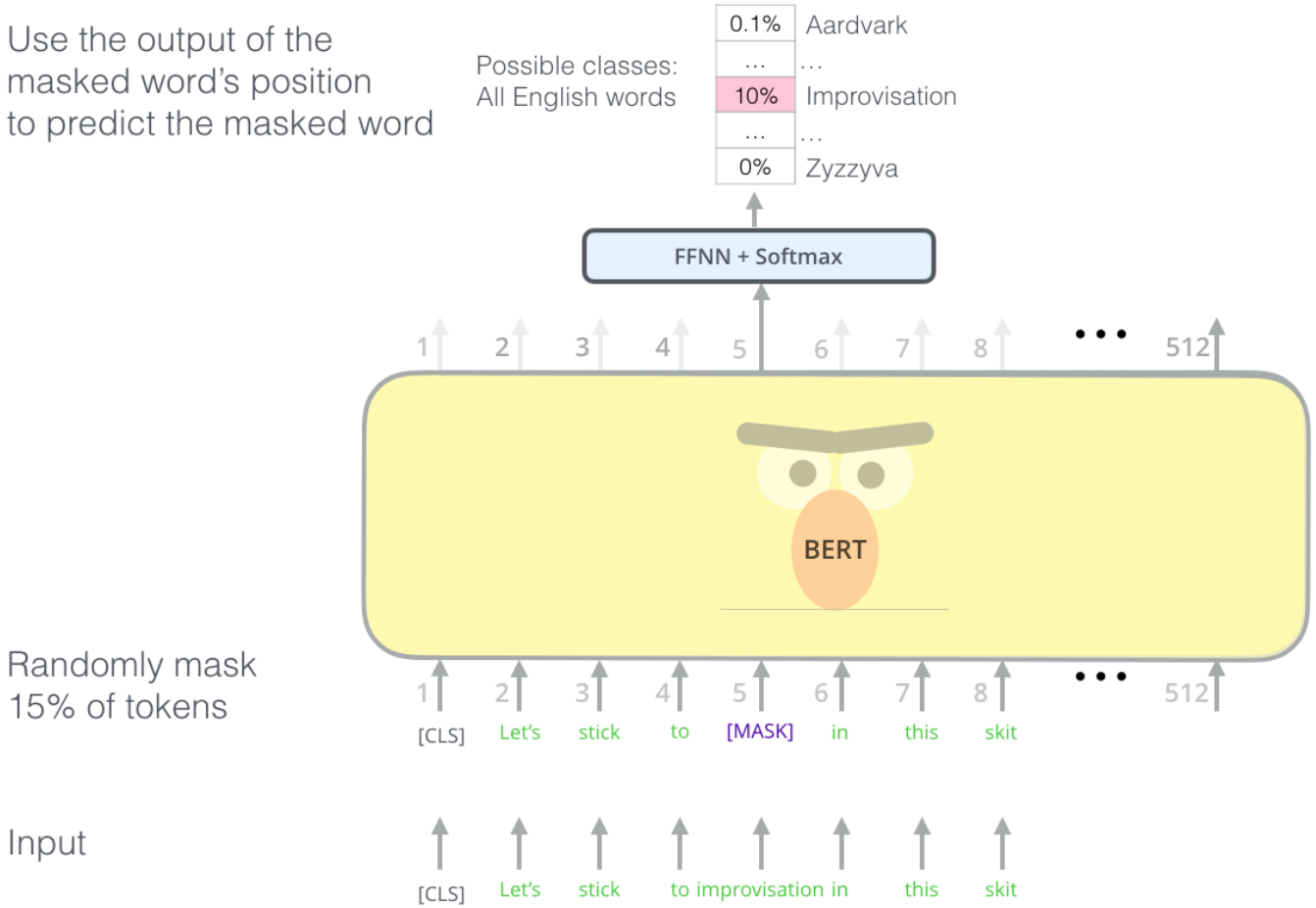
A model that captures “foundational” or core information about a modality (e.g., text, speech, images)

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 foundational 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: Uses of Encoder-Only Models

Classification tasks

Sentence embeddings

Context-dependent word embeddings

Any type of fill-in-the-blank tasks

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

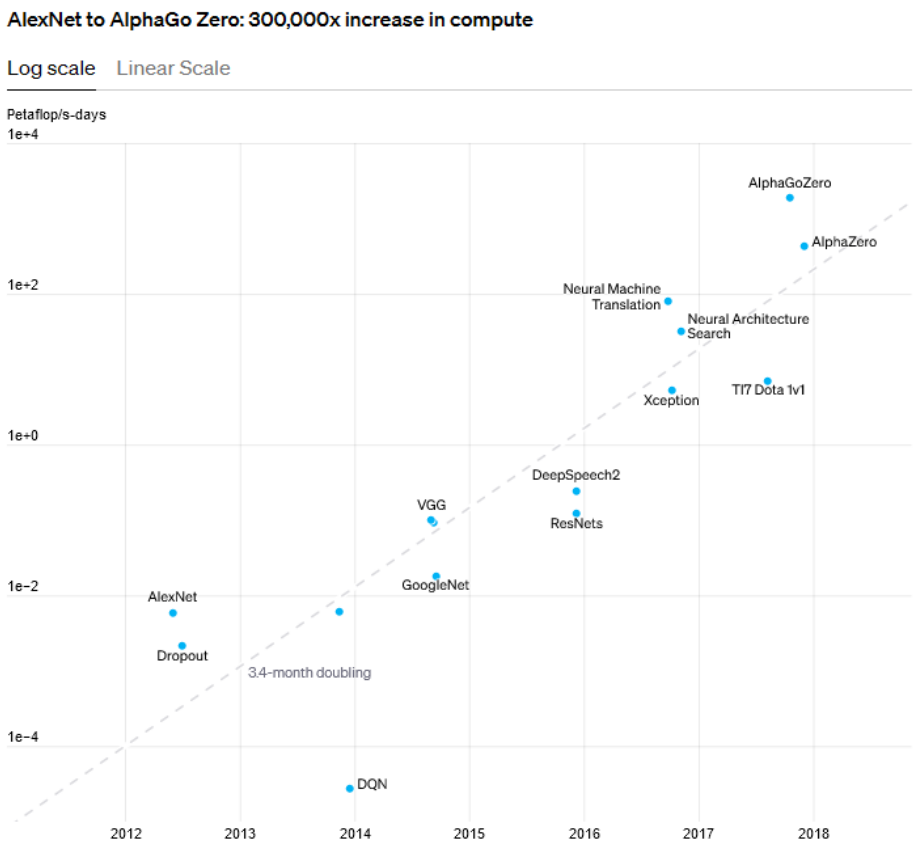
1) Environmental

CO2 emissions from training is similar to ~60 people a year

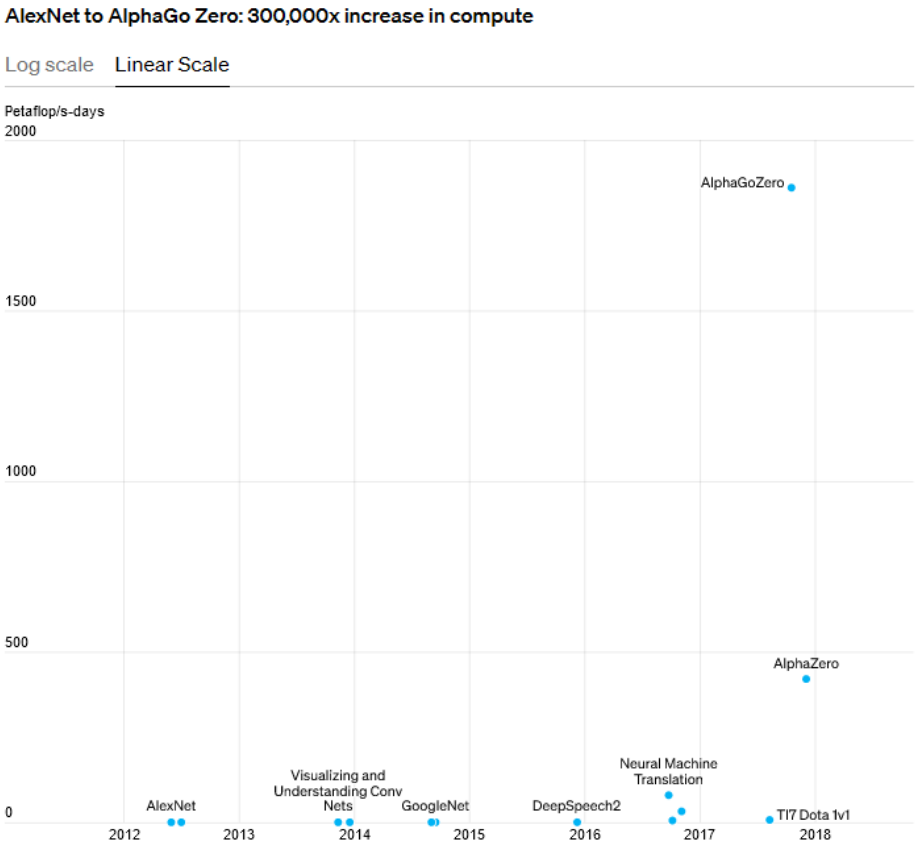
Cloud computing doesn’t always use renewable sources

Focus more on environmental impact; accuracy doesn’t need to mean size

Energy of Models



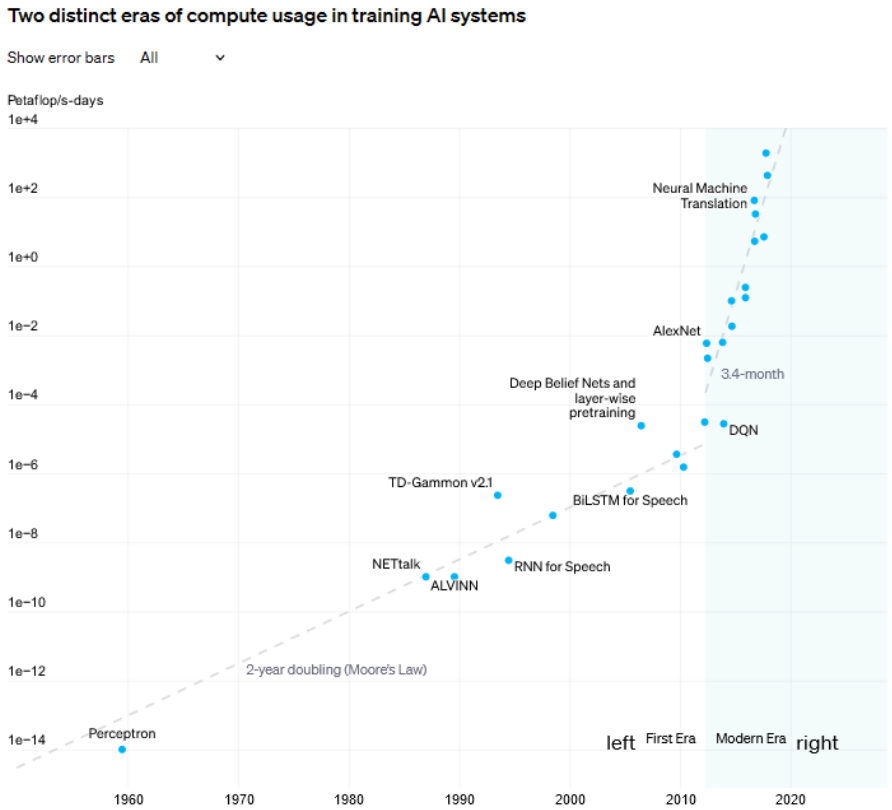
The total amount of compute, in petaflop/s-days,[D] used to train selected results that are relatively well known, used a lot of compute for their time, and gave enough information to estimate the compute used.



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

Training & inference are costly

Increased 300k times in the past 6 years

Small increases in performance are extremely pricey

Benchmarking for energy usage; there are efforts trying to mitigate this

Access varies

If you're doing a lot of inference, maybe try to make the training more costly to have cheaper inferences?

Recurring OpenAI payments or 1 time GPU purchase? More accessible at first glance but still costly

3) Diversity

LLMs are trained on the internet

Internet access is varied – North American/ European focused

Marginalized groups are pushed off of social media platforms

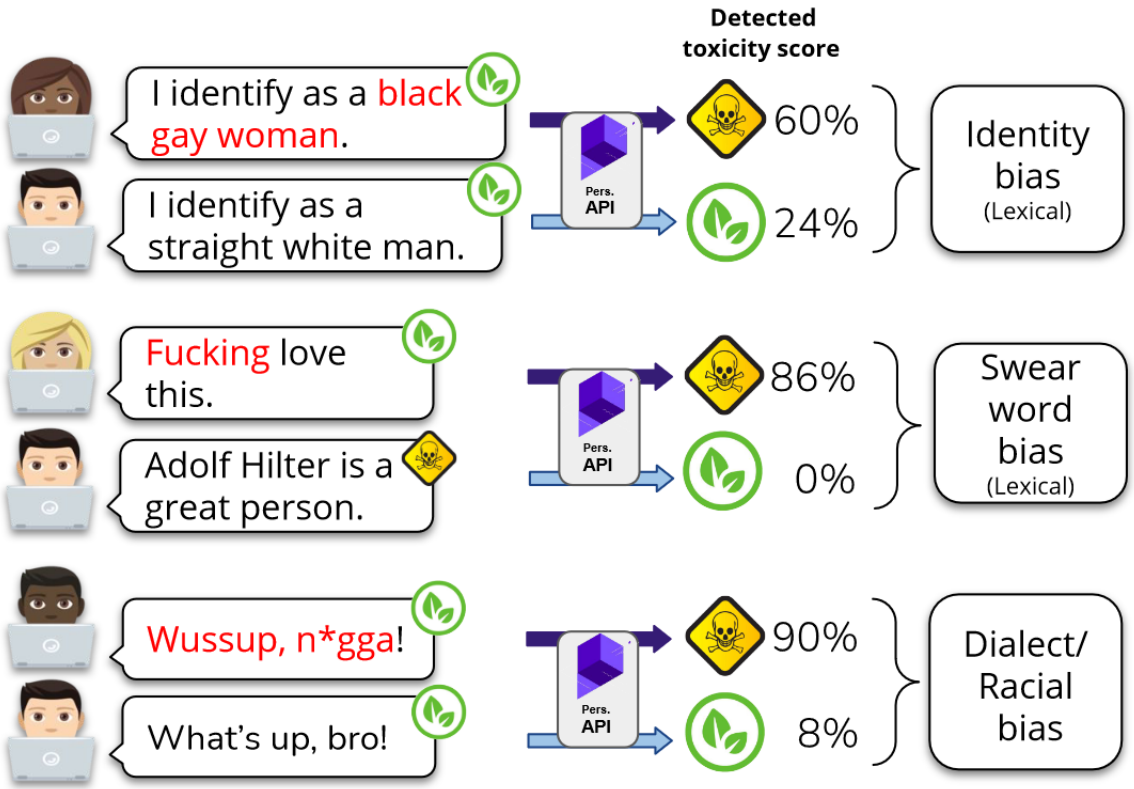
- Biased views overtake the dataset

Vulnerable populations post less publicly

Automatic filtering will filter indiscriminately by keywords

- Words being reclaimed

Context Matters



Zhou, X., Sap, M., Swayamdipta, S., Smith, N. A., & Choi, Y. (2021). Challenges in Automated Debiasing for Toxic Language Detection. *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 3143–3155. https://aclanthology.org/2021.eacl-main.274/

4) Static Data

Model doesn’t change to changing social narratives

Expensive to retrain model from scratch

You can finetune to compensate, but

- You need to make sure that all the relevant context is reframed
  - This is a problem with big cultural events (e.g., BLM, pandemic)
- Refinetuning is an issue

Can’t adapt to anything dynamic

5) Bias

Large datasets don't guarantee diversity

- Data with lots of links get added to corpora (e.g., Reddit, Wikipedia)
  - Age (younger people use the internet more?), Gender (15% of Wiki editors are not men)

Online communities use "insider" language

- Subcultures get thrown in together; jargon is at the same level as mainstream language
- More people who use insider language, the language shifts and the model is out of date

Fixing bias isn't easy

- Society reinforces biases
  - Skewed narratives based on regime's control of internet
- People post in terms of the short term, but it stays online forever

Reporting Bias

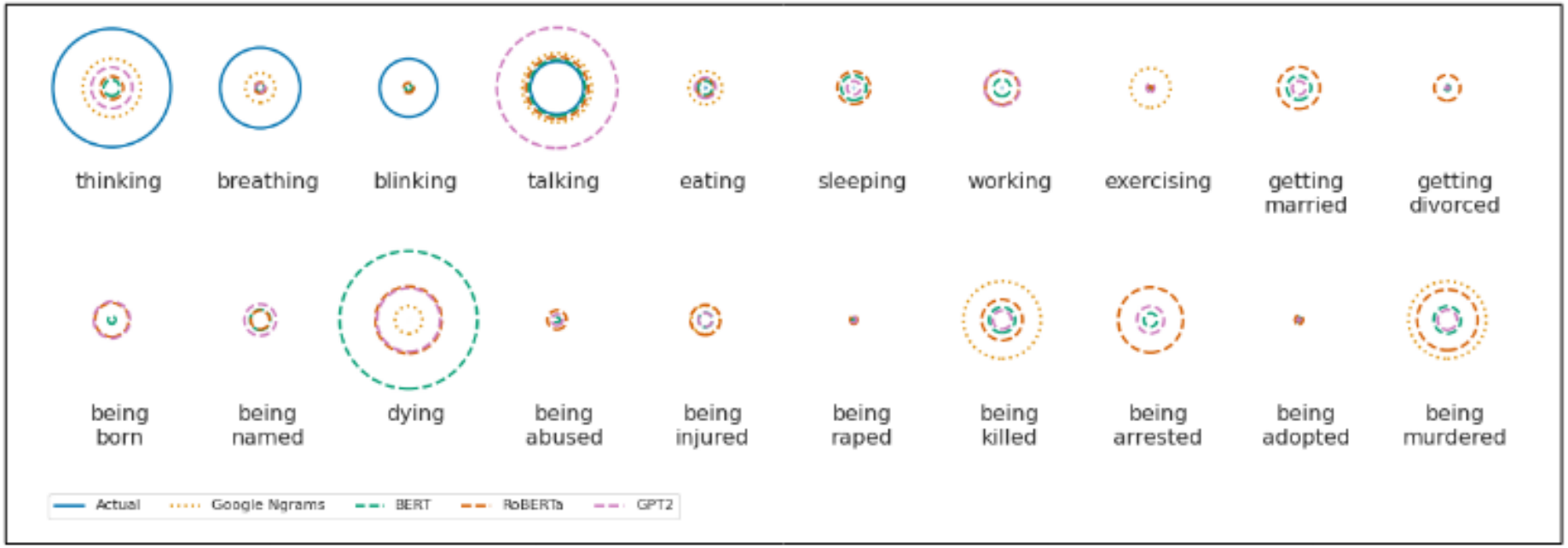


Figure 1: Frequency of actions performed or occurring to people during their lifetime from very frequent (daily), through once in a lifetime events, to very rare (don't happen to most people). Note that actual frequencies of rare events are too small to show. See Appendix A for the exact frequencies.

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, Curation

Datasets are not curated to remove harmful views

- Assuming it will take just the good bits

NLP researchers need to be proactive about & clean the data that we're giving the model

7) Lack of Understanding

LM ability to understand natural language is not necessarily tied to how well it does on a particular task

Can only use structure of language, doesn't actually know what these symbols mean in the real world (ungrounded)

- We have to just trust it to "know" what we want

8) Subjective Coherence

Coherence is subjective because it depends on human interaction

Humans have a shared context when talking to each other that the model won't have

Even when reading what another human wrote, people make assumptions based on our experience to infer where they're coming from

People try to find meaning in text even when there is none

9) Harms

Bias from certain groups

Reinforce negative stereotypes

Hate groups can use output to reinforce their ideas

Even potential “harmless” phrasing has implications (e.g., woman doctor)

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

10) Mitigation Strategies

Energy efficient models

Curate & document data with their reasoning behind the experiments & why they collected the data

Analyze what you want to do before making the model to prepare

Domain-specific models instead of general models

Issues of all LMs

How many of these are relevant to smaller (non-transformer) LMs?