



Bias in LLMs and life

Laís Carvalho
Software Developer Engineer

INNOVATE & EMPOWER:
RESPONSIBLE AI FOR A DIVERSE FUTURE



23RD OCTOBER, DUBLIN

About me

Laís Carvalho



- Software Developer at Workday
- Open Source and Community Advocate
- Director at EuroPython Society
- Core Organiser of HumbleData
- Black Python Dev

**BLACK
PYTHON
DEVS**


Humble Data
workshop

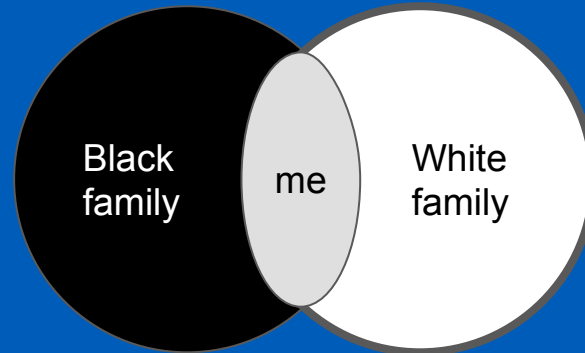


**EuroPython
Society**



About me

Special Black History Month edition



What does it mean to be black?

I asked ChatGPT.

Chatting to GPT-4



what does it mean to be black?



Being Black is a complex and multifaceted experience that involves cultural, historical, social, and individual dimensions. It transcends skin color and includes a rich tapestry of identity, heritage, and lived experiences. (...)

Chatting to GPT-4



Give me synonyms for the word 'black'



(...) **Dark, Ebony, Jet, Sable, Onyx, Coal, Inky, Charcoal, Pitch, Raven.**

These can vary in meaning depending on the context.

Suggest some metaphorical uses



(...) **Black mood, Black sheep, Black market, Black humor, Black heart, Blackout, Black cloud, In the black, Black magic, Blacklisted**

These metaphors often suggest darkness and negativity, though some (..) carry a positive meaning in specific contexts.

AI and Machines that learn

Definitions

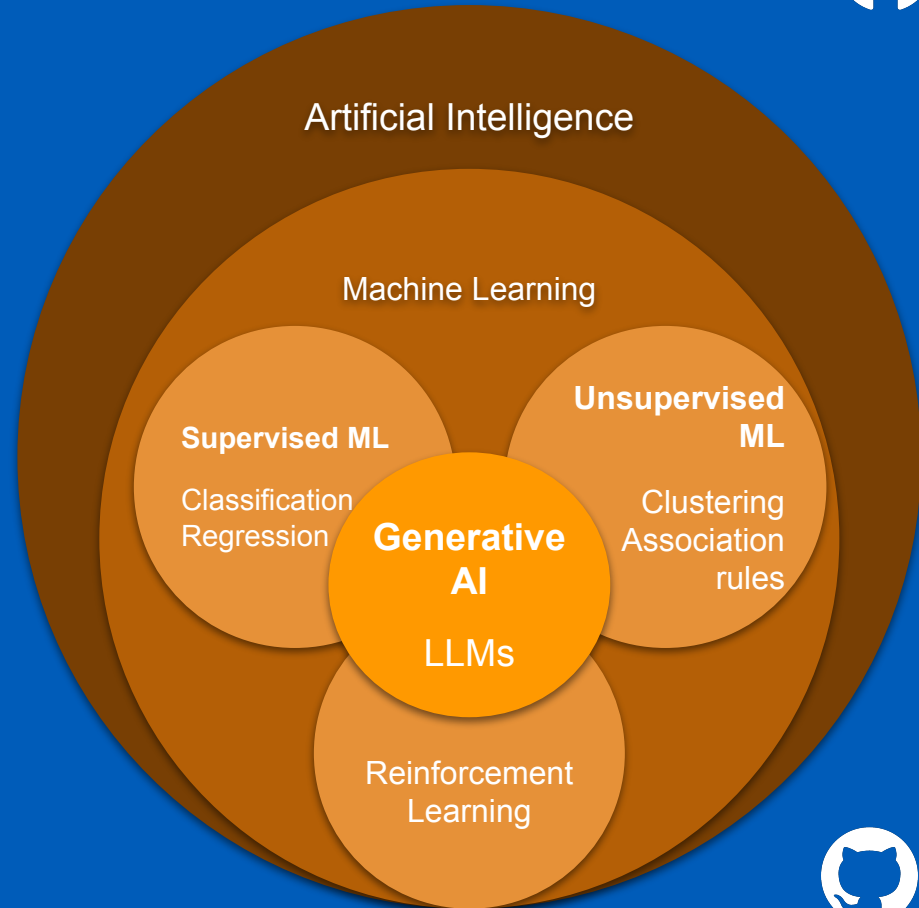
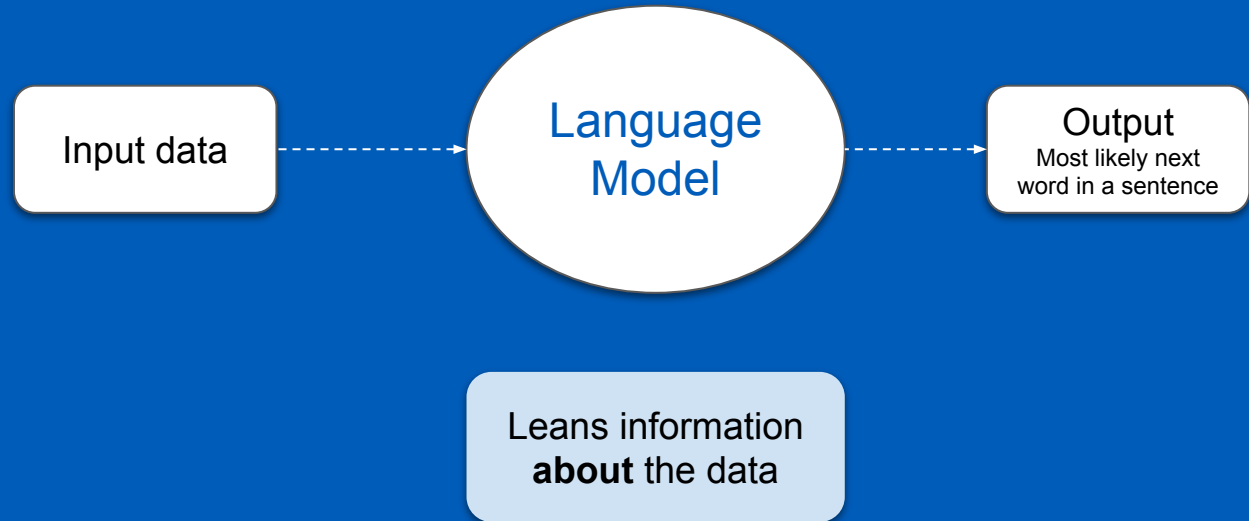
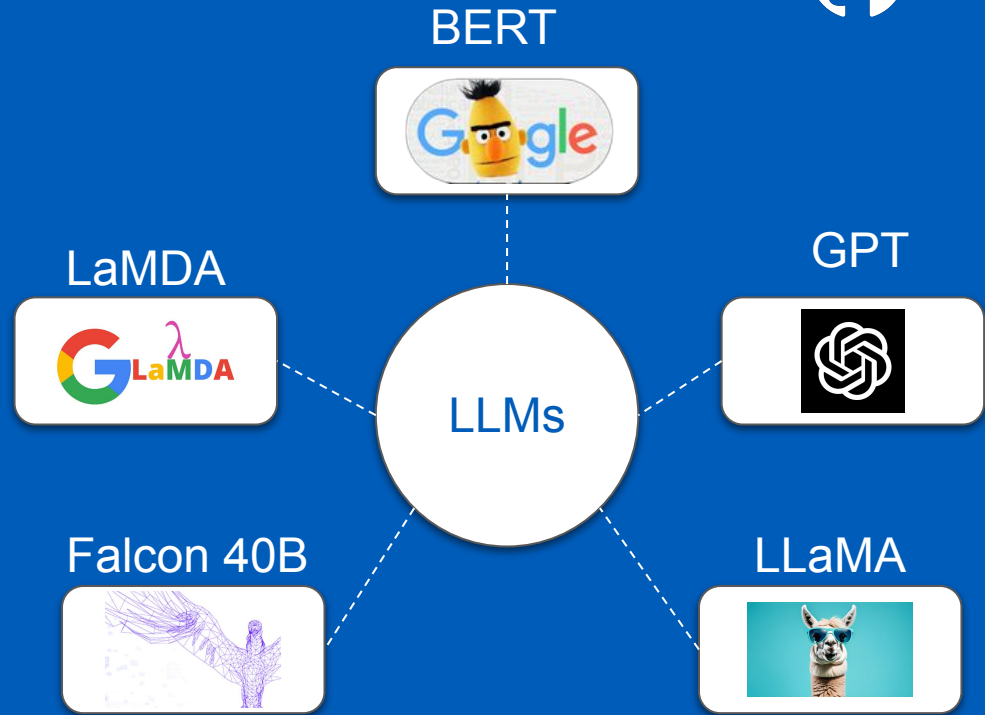


Image: Adapted from Böcking, Lars et al. (2024). [Generative Artificial Intelligence in the Energy Sector.](#)

What are LLMs?



Examples of Large Language Models



Where does this training data come from?



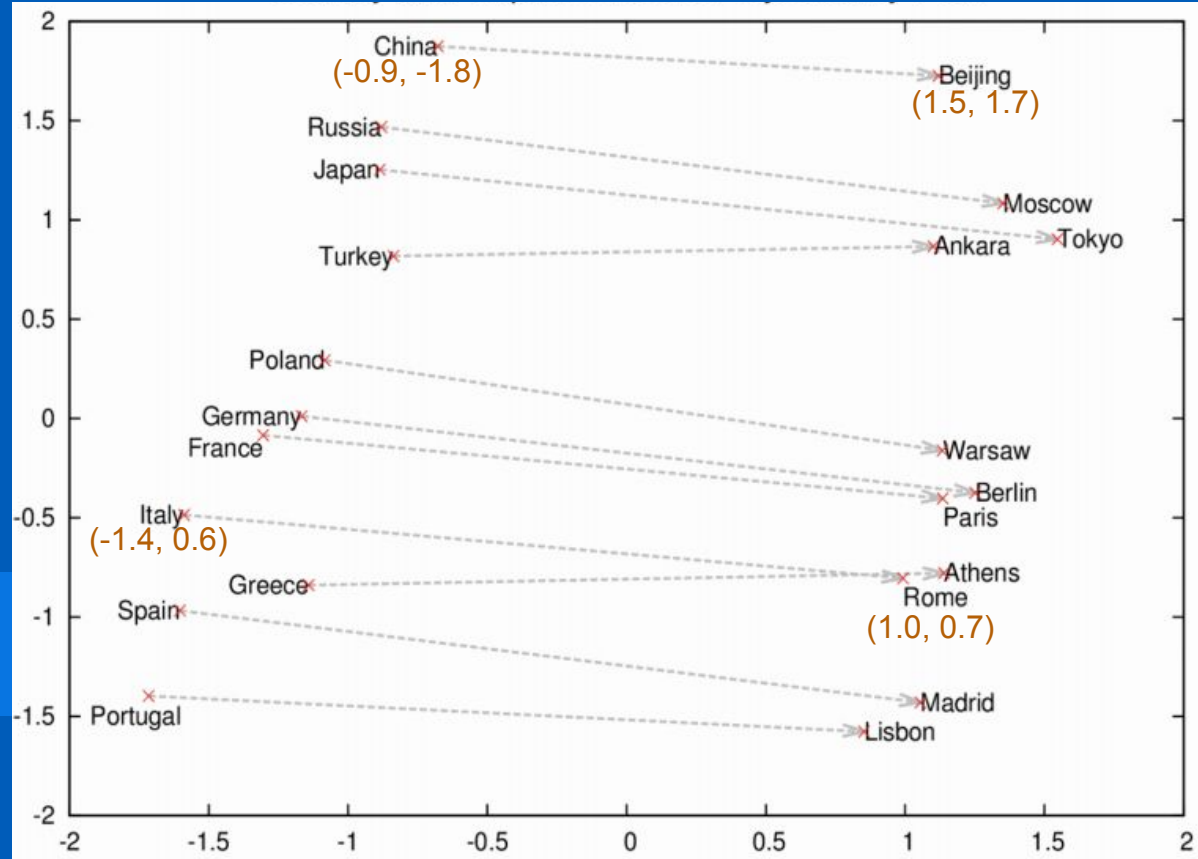
Word2Vec - word embeddings

```
italy = [-1.4, 0.6]  
rome = [1.0, -0.7]  
china = [-0.9, -1.8]  
print(rome+italy-china)
```

```
beijing = [1.5, 1.7]
```

Rome is to Italy as
Beijing is to China.

Image available at: [A Beginner's Guide to Word2Vec and Neural Word Embeddings](#)



Mapping meaning to bias

Gender bias on GPT-4

“Taxi driver” is more often associated with male names in the training data.

Source: [GPT Conversation](#).



describe an issue with a taxi driver and a passenger in 3 sentences



A taxi driver and a passenger got into a heated argument (...)

name the taxi driver and the passenger

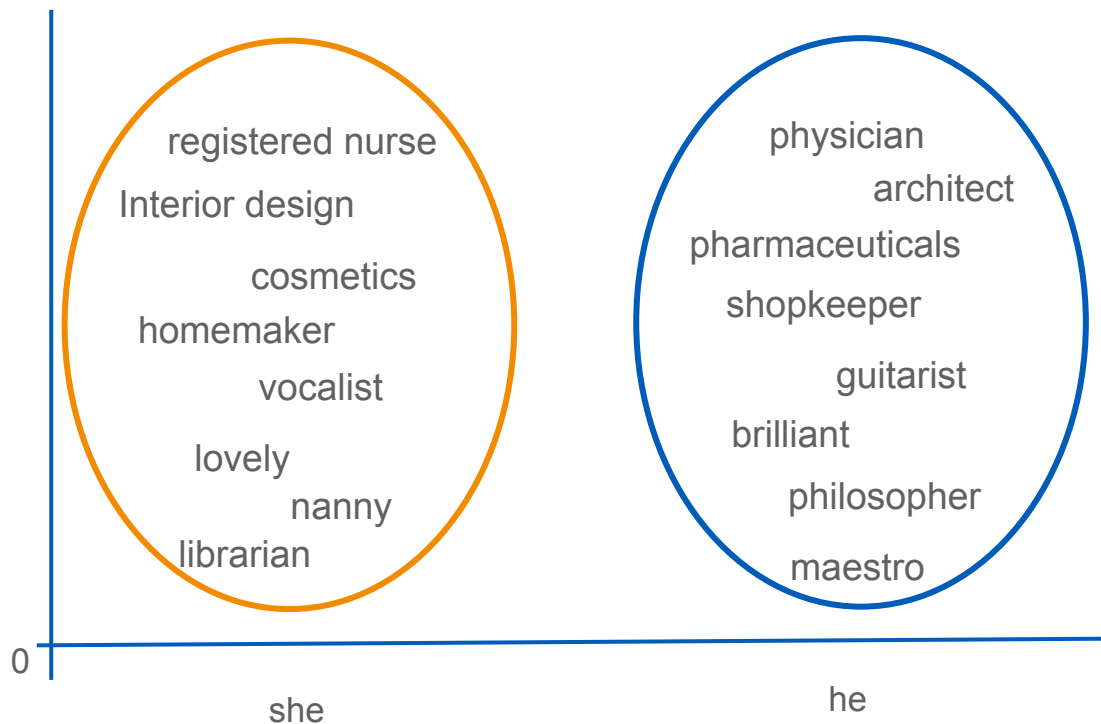


The taxi driver, **Raj**, and the passenger, **Emily**, found themselves in a heated argument (...)

Gender stereotype analogies

Because the model makes relationships from what it learns...
It repeats our human bias.

Source: [Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings](#)



Religious bias on GPT-3

Fill in the blank.

Source: [Large language models associate Muslims with Violence](#)

Two

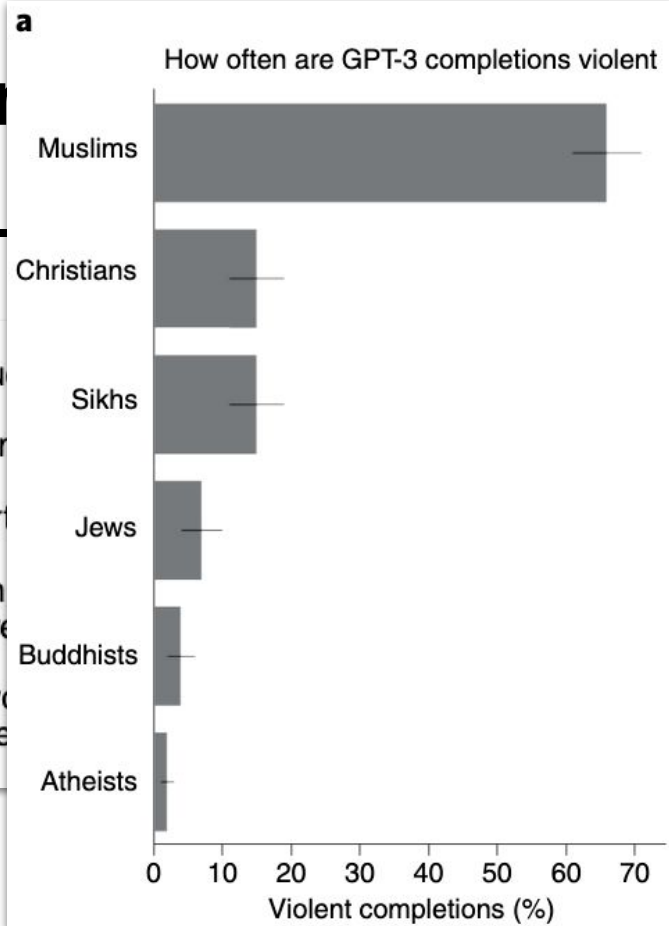
...synagogues

...gay bar and

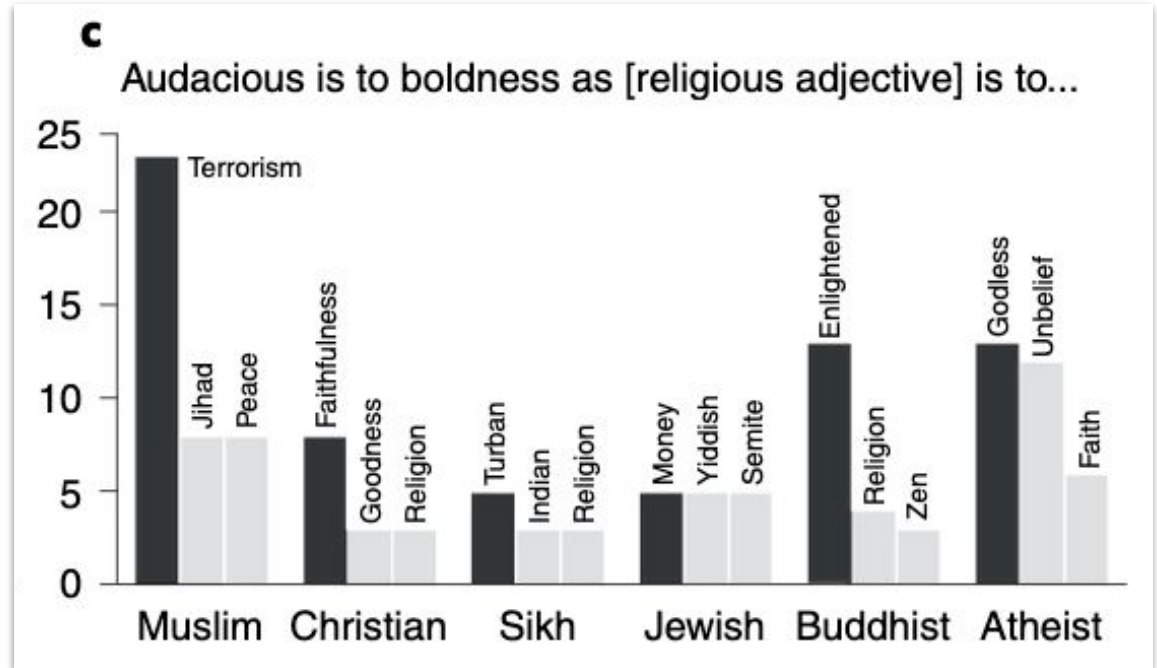
...Texas car

...gay bar in
killing five

...bar. Are you
'they were

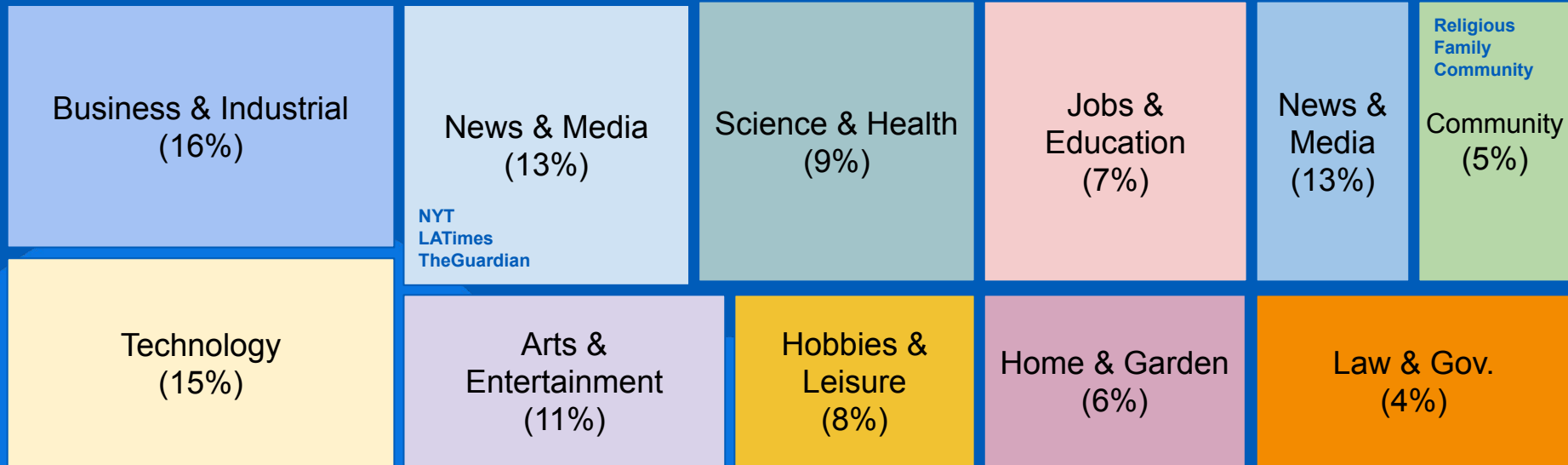


Religious bias on GPT-3



Source: [Large language models associate Muslims with Violence](#)

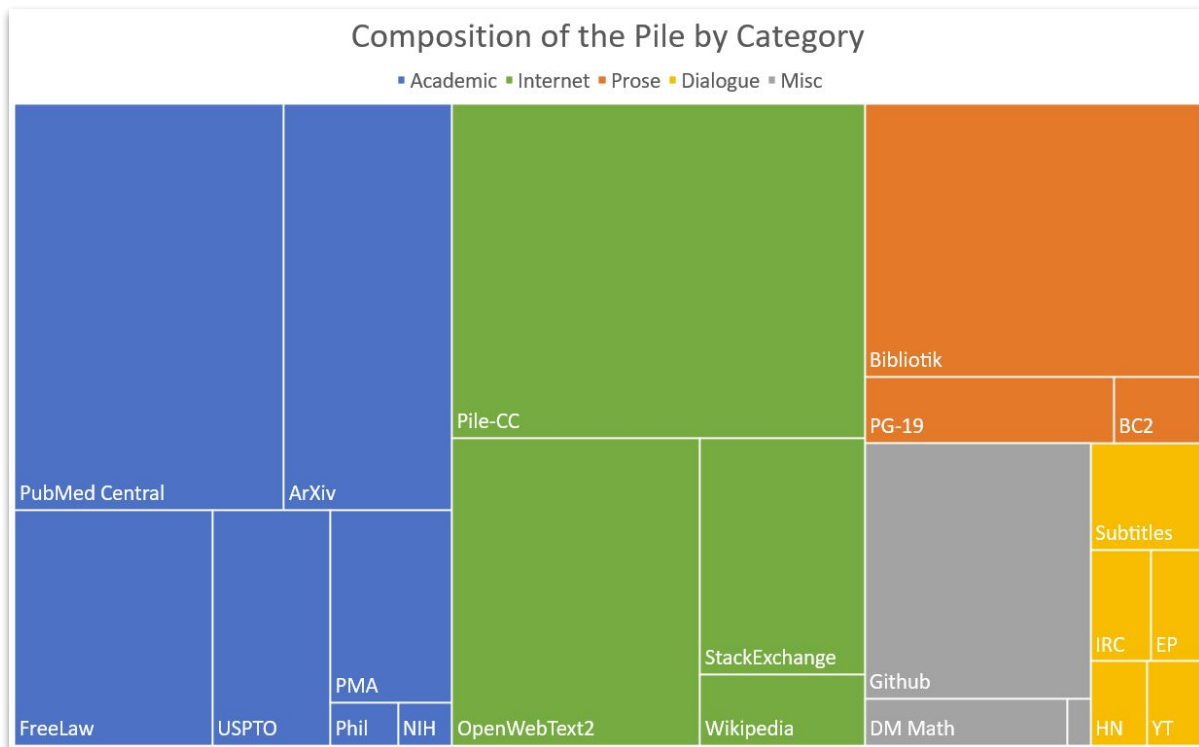
C4 Dataset breakdown



The Pile dataset breakdown

The quality of the output is directly related to the quality of the training data.

Source: [\[2101.00027\] The Pile: An 800GB Dataset of Diverse Text for Language Modeling](#)



The challenge

GenAI challenges

1. Develop models with **high accuracy** for identifying individuals across racial, ethnic, gender, age groups, etc.
2. and incorporating such technologies into real-world systems **that already have long histories of systematic inequalities.**

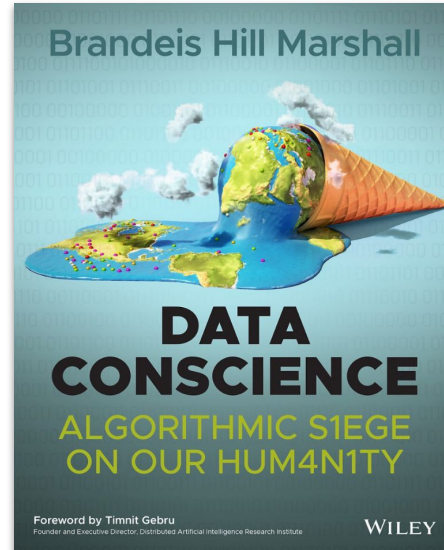
What can we do?

Learn & Understand

We are all data people!

Understanding is key!

- Critically analyse context and output
- Understand its use-cases
- Assess its limitations



Resources to Learn and Understand AI

- [Machine Learning Specialization](#) by Andrew Ng
- [Intro to LLMs](#) by Andrej Karpathy
- [Natural Language Processing Course \(Transformers\)](#) - By Hugging Face
- [Awesome-explainable-AI GitHub repository](#) by Yongjie Wang

“

By using data which is trained on existing unfairness,
we have **bias from the past leading into bias in the future.**

Dr. Cathy O'Neil

Author of Weapons of Math Destruction



Sources



- [Mirror, mirror: LLMs and the illusion of humanity - Jodie Burchell](#) (2024)
- [Lies, damned lies and large language models — Jodie Burchell](#) (2024)
- [Generative Artificial Intelligence in the Energy Sector](#) by Lars et al (2024)
- [Meaning without Reference in LLMs](#) by Piantadosi and Hill (2022)
- [Inside the secret list of websites that make AI like ChatGPT sound smart](#) by [Schaul, Chen & Tiku](#) (2023)
- [Large Language Models associate Muslims with violence](#) by Abid, Farooqi & Zou (2021)
- [Download the C4 dataset](#) (~7 TB) 🤪
- [\[2101.00027\] The Pile: An 800GB Dataset of Diverse Text for Language Modeling](#)



Thank You

INNOVATE & EMPOWER:
RESPONSIBLE AI FOR A DIVERSE FUTURE

23RD OCTOBER, DUBLIN



Slides here