

CS 1671/2071

Human Language Technologies

Session 25: Social factors, bias and ethics in NLP

Michael Miller Yoder

April 16, 2025

Course logistics

- Next class on Mon Apr 21 will be in-class project work time
 - Michael will be available to answer questions
- No class after that until final presentations **Wed Apr 30, 12-1:50pm**

Course logistics

- Final report is **due Mon Apr 28**
 - Instructions are released on the project website
 - Maximum 8-page report in ACL format (Word and LaTeX templates [here](#))
 - Abstract, introduction, data, methods, results, discussion, future work, limitations, ethical issues, group member task breakdown, references, appendices (optional)

NLP seminar by Jieyu Zhao today

- Virtual talk on Zoom: today, Apr 16, 4:30-5:30pm
- <https://pitt.zoom.us/j/95500691262>
- Trustworthy Language Models
- Understanding and eliminating unwanted behaviors, including auditing NLP models, detecting and mitigating biases, and understanding how LLMs make decisions



Overview

- Language in social context
- Computational social science
- Bias and ethics in NLP
 - *Warning: slides contain offensive stereotypes*

Language is embedded in social context

What types of social situations do you encounter language in?

What types of social contexts?













English chatting very easy



Alan Black
has your back.

DWI
Criminal Defense
Family Law
Personal Injury

What types of social contexts?

Euro quals	2:45 PM ET ESPN3  BEL  CYP	2:45 PM ET ESPN3  NED  EST	2:45 PM ET ESPN2/ESPN3  WAL  HUN	2:45 PM ET ESPN3  GER  NIR	2:45 PM ET ESPN3  POL  SVN
------------	--	--	--	--	--

On-Line Homework Instructions for Physics 1250-1251

Homework will be submitted and graded via the online software package WebAssign.

ACCESSING WEBASSIGN:







Open Internet Explorer or Netscape Navigator or Mozilla Firefox (Some other browsers may have difficulty), and go to the WebAssign login page (<https://www.webassign.net/osu/student.html>). (The WebAssign login page at <https://www.webassign.net/login.html> will get you to the site above as well, but the OSU login site should be your primary site.)


What types of social contexts?




What types of social contexts?

What's happening?

      [Tweet](#)

 **Odd Pittsburgh** @OddPittsburgh · 59m
[#Pittsburgh](#) in 1930



City of Pittsburgh

Trends for you

- Trending in United States
#DevinNunesIsAnIdiot
53.9K Tweets
- Trending in United States
#AdviceForBoomers
4,684 Tweets
- Trending in United States
Vindman
Trending with: Lt Col Vindman, Colonel Vindman, Col Vindman
- Trending in United States
#2009v2019
3,035 Tweets

[Show more](#)

What types of social contexts?



World

U.S.

Politics

N.Y.

Business

Opinion

Tech

Science

Health

Sports

Arts

Books

Style

How Not to Plot Secret Foreign Policy: On a Cellphone and WhatsApp

U.S. officials expressed wonderment that Rudy Giuliani ran an “irregular channel” of Ukraine diplomacy over open cell lines and apps penetrated by the Russians.

2h ago [494 comments](#)



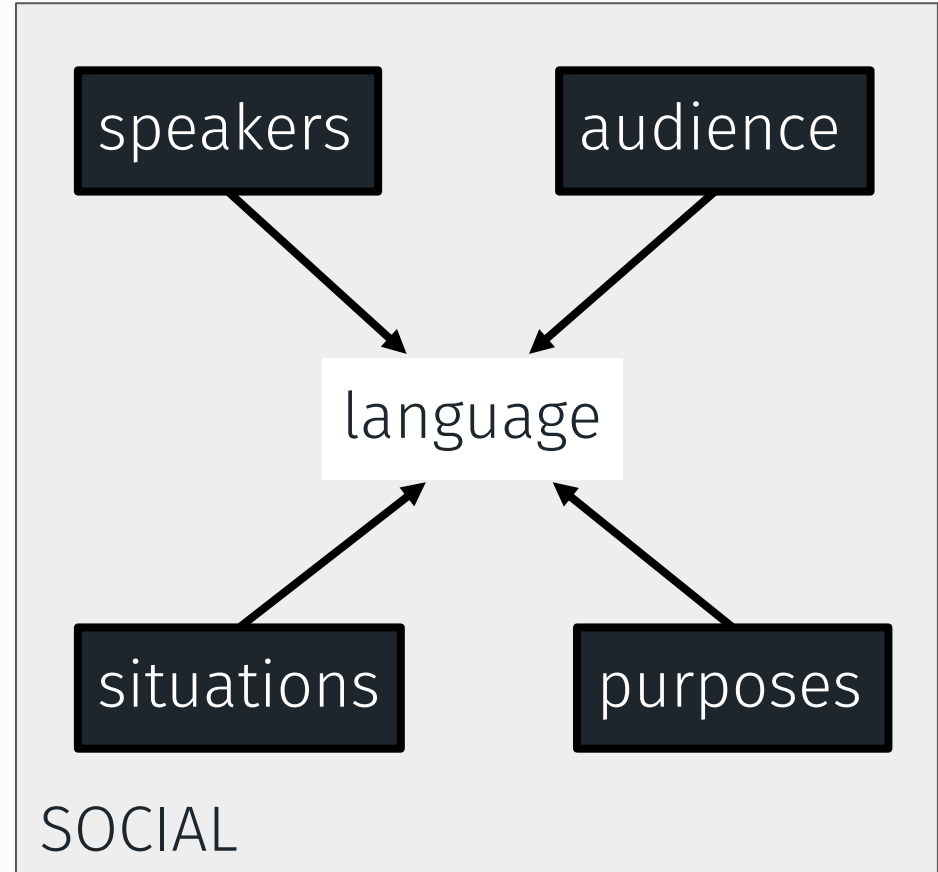
Rudolph W. Giuliani, President Trump's personal lawyer, makes a living selling cybersecurity advice.
Doug Mills/The New York Times

Who is Kurt Volker, President Trump's former special envoy to Ukraine?

28m ago

Tim Morrison, a hawkish aide loyal to Mr. Trump, will also testify this afternoon.

42m ago



NLP + social science: applications

hate speech detection

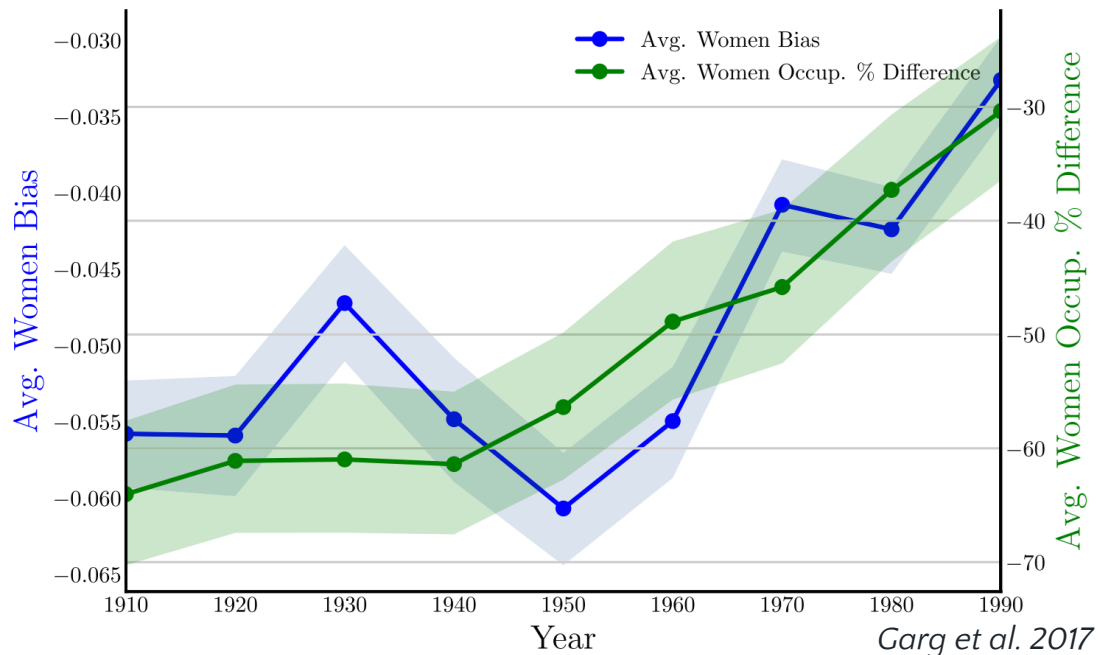


community norms



NLP + social science: applications

fairness and bias

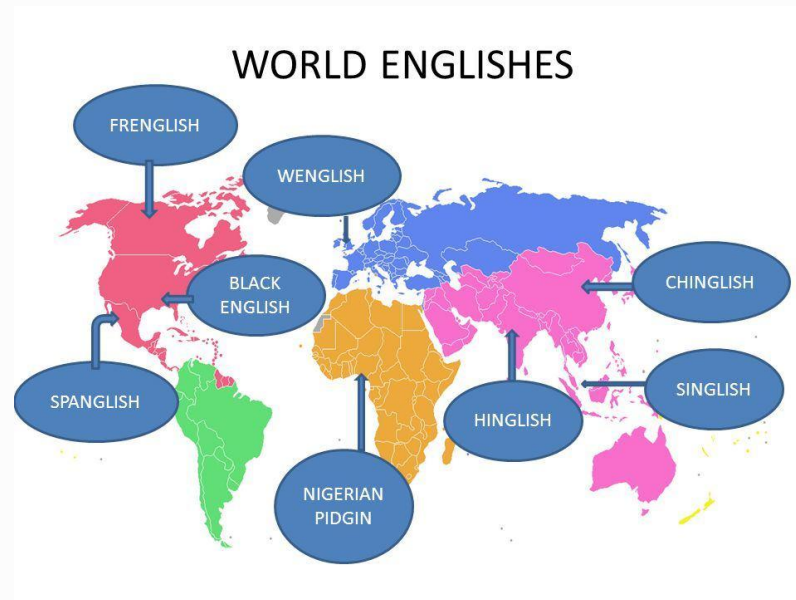


media framing



<https://criticalmediareview.wordpress.com/2015/10/19/what-is-media-framing/>

dialectal NLP tools



Computational social science

Computational social science

- Investigating (modeling, analyzing) social phenomena with computational tools [Cioffi-Revilla 2017]
- CSS goal: find out something about **people** (social science)
- NLP goal: build computational tools that can process or produce language
- CSS+NLP: using NLP tools to measure or predict social information from language use

Computational social science: methods and data

- Observational studies, not lab or survey studies



large datasets of social interaction

Computational social science example

Example: How fast does fake news spread? [Vosoughi et al. 2018]

$$y = f(x)$$

spread through a network

network analysis

true/fake news

NLP/text mining

Computational social science example

Example: Do police officers speak more respectfully to white drivers than Black drivers in traffic stops? [Voigt et al. 2017]



● Bias and ethics in NLP

Bias and ethics in NLP [Hovy and Spruit 2016]

- Language, society and individual are interrelated
 - We must think about ethics when dealing with people
- Demographic bias: language contains latent information about the people who produced it
 - Exclusion of the language of people not represented in training data
 - Bias toward Indo-European languages and a few “high-resource” languages in NLP
- Dual use: reinforce prescriptive linguistic norms and degrade non-standard language use with educational language technologies “correcting” language
 - NLP can detect fake news, but also generate it
 - Authorship attribution could identify political dissenters
- Funding sources for research include military. Whose interests are embedded in systems?

The “Bender Rule” [Bender 2019]

- When doing NLP work, please **name** the languages you are working with
 - “Always name the language(s) you’re working on”
- Don’t just assume the “default” language is English and work on other languages is “language-specific”
- English has particularities
 - Massive amounts of training data available
 - Relatively fixed word order
 - Few inflectional forms per word (not much morphology)
 - Orthography: words indicated by whitespace, roughly phone-based

Discussion

- What are some language technologies that you see as particularly needed for a language other than English that you are familiar with? For example, better machine translation, Internet search, speech transcription or recognition, etc.
- What are possible harms of having English be the “default” language for NLP research and system development?

LLMs are powerful, but there are ethical concerns

- Bias & Discrimination
- Hate Speech & Toxicity
- Misinformation
- Privacy

Allocational harms [Crawford 2017]

System unjustly allocates or withholds opportunities or resources to groups




Source: Gizmodo




Representational (associative) harms [Blodgett et al. 2020]

System reinforces subordination or stereotypes about groups

He is...



She is...




Gender bias in word embeddings
[Bolukbasi et al. 2016]

The New York Times

'Nerd,' 'Nonsmoker,' 'Wrongdoer': How Might A.I. Label You?

ImageNet Roulette, a digital art project and viral selfie app, exposes how biases have crept into the artificial-intelligence technologies changing our lives.



Bias & discrimination (representational harms)

Neural models encode social biases against marginalized identities

- Case study: language model generation.
- GPT-2 generates text with more negative associations of black, woman, and gay demographics on 'occupation' related topics.

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

Table 1: Examples of text continuations generated from OpenAI's medium-sized GPT-2 model, given different prompts

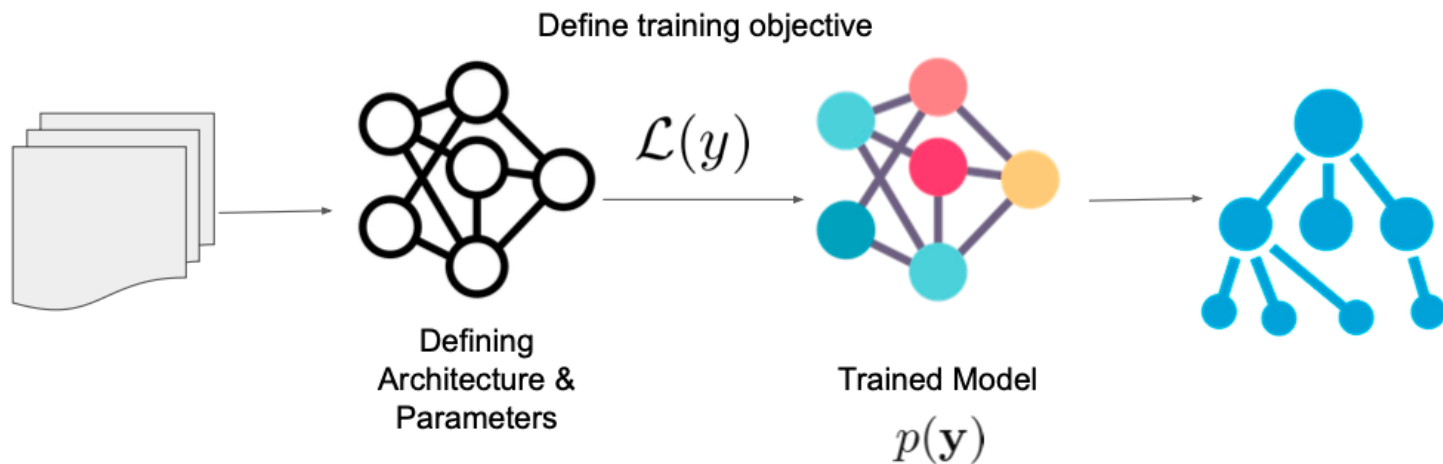
Causes of social harms from LLMs

Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.
– Ruha Benjamin, *Race after Technology*

- Language models were designed to model probability distributions of text. They
 - Do not understand social norms and morals
 - Reinforce and amplify biases
- Uncurated sources of training data
 - Reddit: 67% of Reddit users in the United States are men, and 64% between ages 18 and 29
 - Wikipedia: only 8.8–15% are from women editors
 - Web data contains conspiracy theories, misinformation, aggressive text

Mitigating harms in NLP: design interventions

LLMs: From data to decoding



Data Collection

Data Intervention

Model and Train

Model Intervention

Decode & Test

Decoding Intervention

Postprocessing

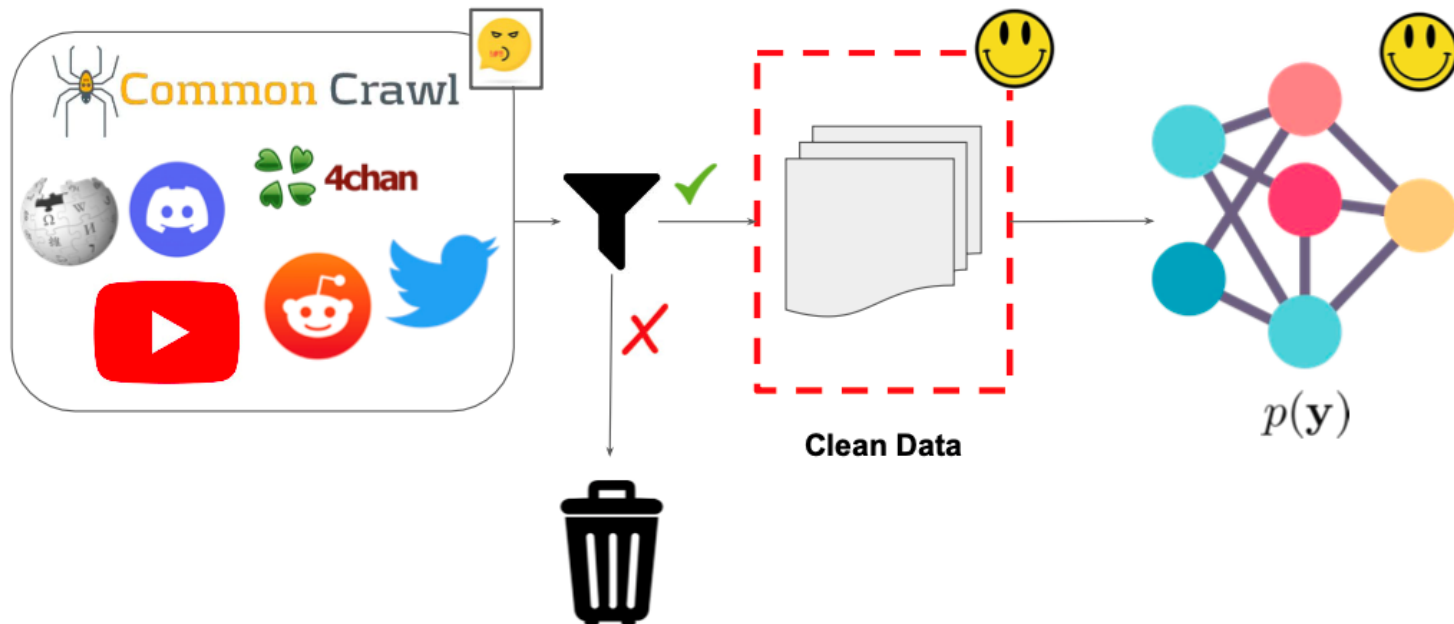
Data statements [Bender & Friedman 2018]

For each dataset you release, document:

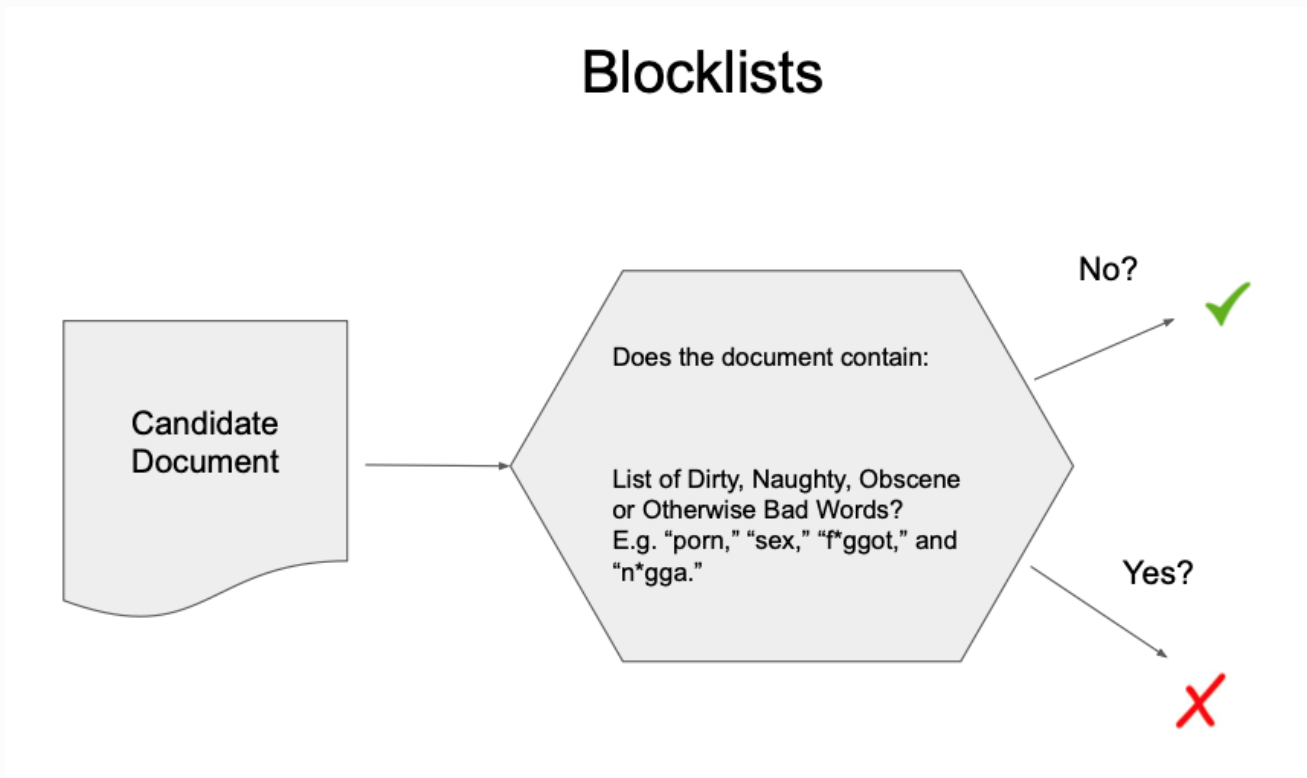
- Curation rationale: why were certain texts selected
- Language variety
- Speaker demographic
- Annotator demographic
- Speech situation
 - Time and place, modality, scripted vs spontaneous, intended audience
- Text characteristics
 - Genre, topic
- Recording quality (for speech)

Data interventions

Data Filtration

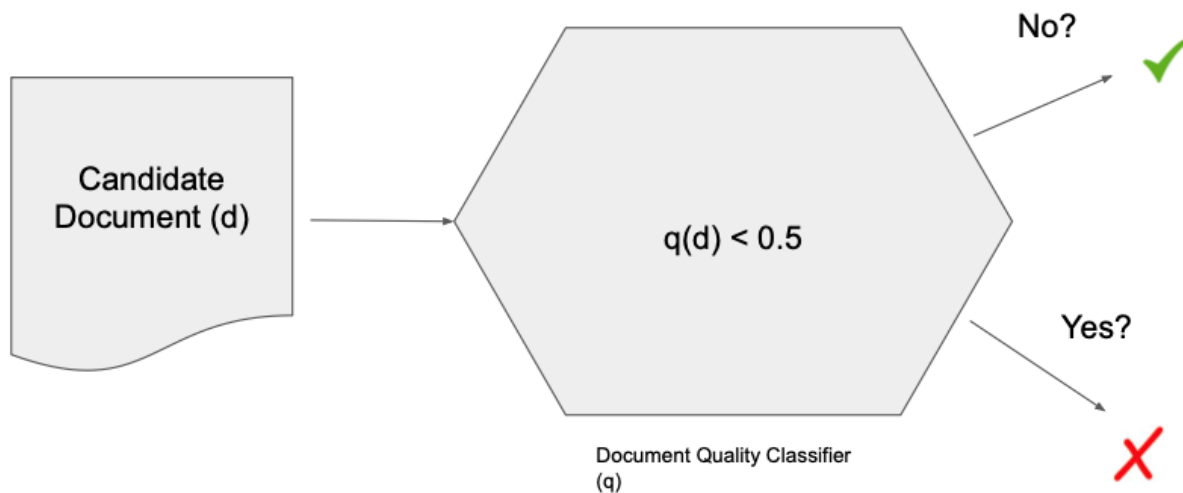


Data interventions: filters options



Data interventions: filters options

Bad Document Classifiers



Data interventions help, but are not enough

Filters themselves have biases

Documents with single presence of “hateful” text are removed.

Subtly harmful text is not captured or filtered.

Minority voices are filtered.

Filtration and retraining is expensive

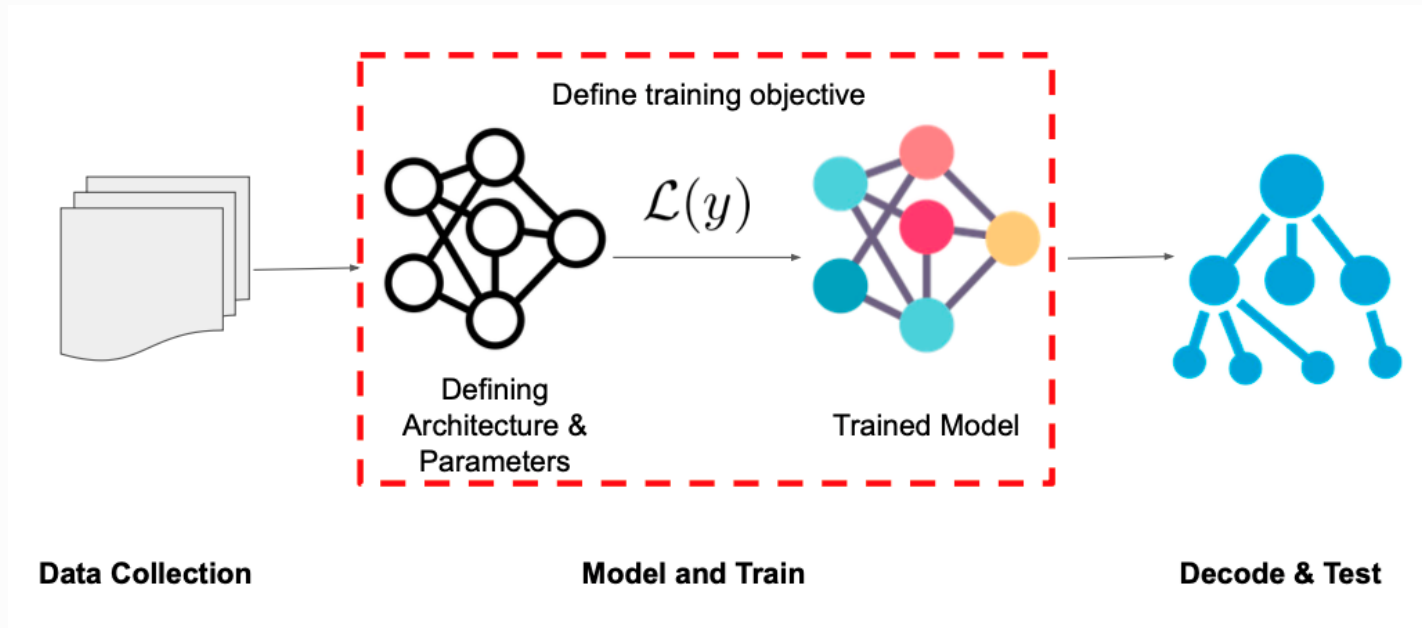
e.g. 175B GPT3 costed an estimated \$12 million to train.

Data is not the only source of issues.

Language models are known to hallucinate information: Lack of **factuality**.

Language models can get outdated and report “false” information.

Model interventions



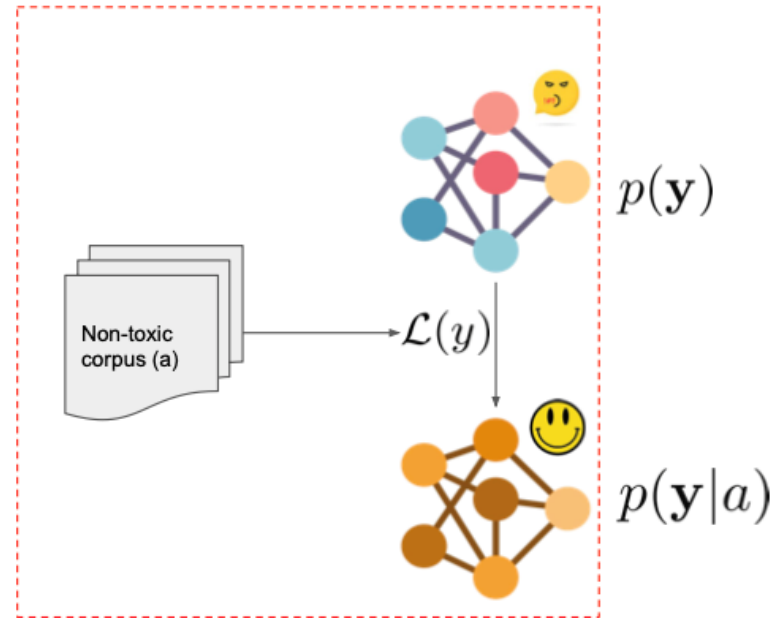
Model cards [Mitchell et al. 2019]

For each algorithm you release, document:

- training algorithms and parameters
- training data sources, motivation, and preprocessing
- evaluation data sources, motivation, and preprocessing
- intended use and users
- model performance across different demographic or other groups and environmental situations

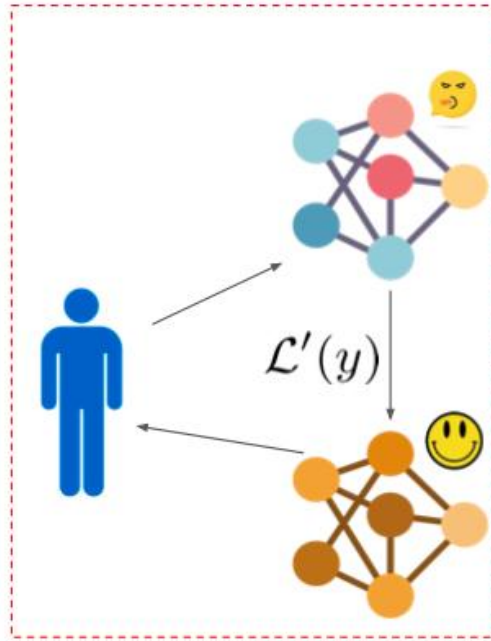
Model interventions: fine-tuning

Modify all model parameters by further training



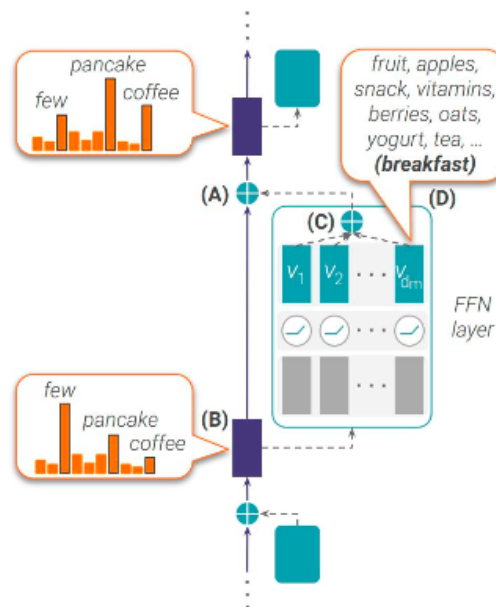
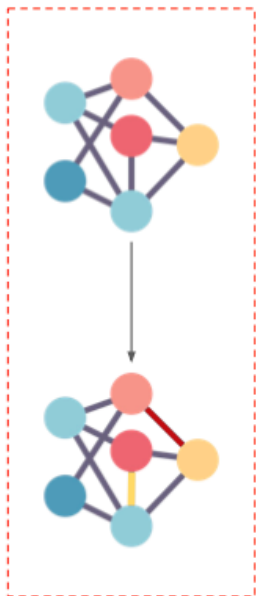
Model interventions: fine-tuning

Modify all model parameters by human feedback



Model interventions: model surgery

Modify specific neurons to “suppress” problematic behavior.



Model interventions help, but are not enough

Architectural modifications

Could require structured data which can be difficult and expensive to collect.

May make training computationally expensive

Surgery

Very context dependent. Difficult to define and remove all problematic behaviors.

Finetuning

Finetuning data is not guaranteed to be good.

Reduces utility of the language models.

Postprocessing

Identify issues with generated outputs and edit them.

Source:

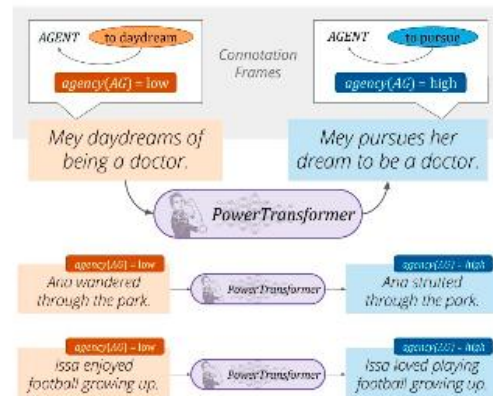
Jerusalem (CNN)The flame of remembrance burns in Jerusalem, and a song of memory haunts Valerie Braham as it never has before. This year, Israel's Memorial Day commemoration is for bereaved family members such as Braham. "Now I truly understand everyone who has lost a loved one," Braham said. (...)

Original: France's memorial day commemoration is for bereaved family members as braham. (inconsistent)

After Correction: Israel's memorial day commemoration is for bereaved family members as braham. (consistent)

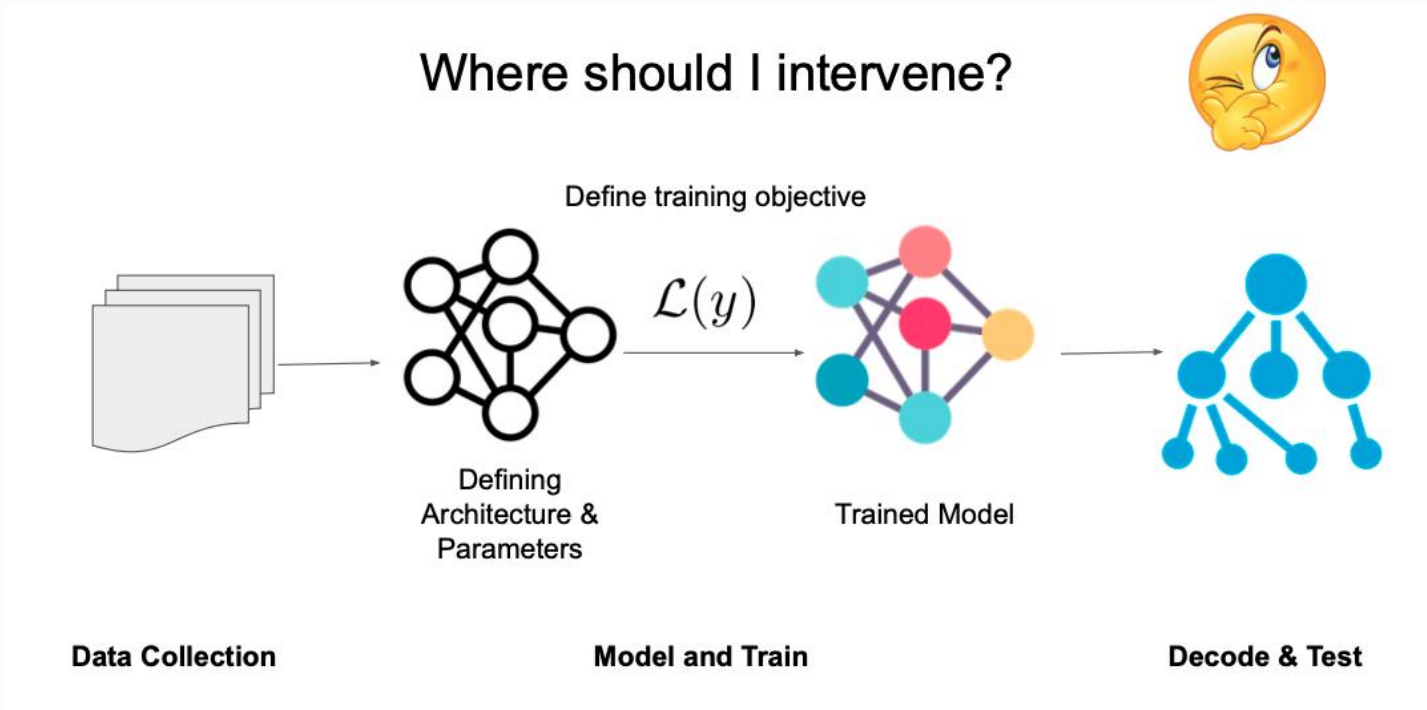
Table 1: An example of an inconsistent system-generated summary and the output summary from our correction model. In this case, "France" is successfully corrected as "Israel".

[Fact correction in summarization, Cao et al 2020]



[PowerTransformer: Debiasing, Ma et al 2021]

Mitigation Summary



Shortcomings of design-based ethics in NLP

- Sometimes the question isn't how to make a system more ethical, but **if it is ethical to build a system at all.**
 - Ask yourself: If language models are the right solution for this problem?
 - Keyes 2019 (tongue-in-cheek): improving fairness across demographics for a system for to “turn the elderly into mulch”
- Most of the presented approaches have potential for misuse.
 - Methods to detect misinformation can be used by adversaries to bypass it.
 - Techniques to reduce toxicity can also be used to make it worse.

Discussion

In an area of academia or industry you are familiar with (in CS or outside CS), do you see issues with transparency of data and/or models?

Is it clear under what circumstances datasets were collected? Are the intended uses of machine learning models clearly stated?

Power and structural issues in NLP ethics

It's about power: ethical concerns of software engineers

[Widder et al. 2023]

- Surveyed 115 software engineers and interviewed 21 software engineers about their ethical concerns, what happens when they develop ethical concerns and **what affects their power to resolve their concerns**
- Military, privacy, advertising, surveillance were top ethical concerns
- Scope of concern: from bugs to questioning entire purpose of an industry
- Refusal, 'quiet quitting' of productivity on a project
- Seeking reassignment to other projects or trying to change the project
 - Pivoting from facial identification to facial verification (i.e. are these the same people?)
- Financial and immigration precarity can make doing something about ethical concerns difficult
- Organizational incentives (making \$\$) that might lead to ethical tradeoff, like selling user data to advertisers when scrambling to find a new revenue stream

Discussion

- Could you relate to any of the ethical concerns raised by software engineers?
- Have you ever had ethical concerns with any work you have done? If not, what types of work might you have ethical concerns about?

Widder et al. 2023 implications

- Not so much about identifying issues, but giving programmers power to address them
 - Ethics checklists and codes can help empower individuals to do this
- Workers need 'guidance on how to navigate organizational power dynamics'
- From a focus on good design -> critique of whose goals are being achieved
- Collective action often needed, but also the role of individuals refusing to work on projects
 - Are they replaceable, as Palantir treated them when employees left over ethical concerns over selling tech to US border enforcement (ICE)?

Language (technology) is power [Blodgett et al. 2020]

- Recommendations for better work on bias in NLP
- Look at fields outside CS for guidance
- Treat representational harms as harmful in their own right
- Explicitly state why “bias” in systems is harmful, in what ways, and to whom. Be explicit about normative reasoning behind these judgements
- Engage with the lived experiences of members of communities affected by NLP systems. Reimagine power relations between technologists and such communities.

Conclusion

- Language is embedded in social context
- Computational social science studies people and societies with computational models of observational data
 - Often uses NLP for analyzing text
- Social biases can be encoded across the NLP system pipeline: data, modeling and use of systems
- Beyond design interventions, consider whose interests NLP systems are serving and push for greater accountability of such systems to all users