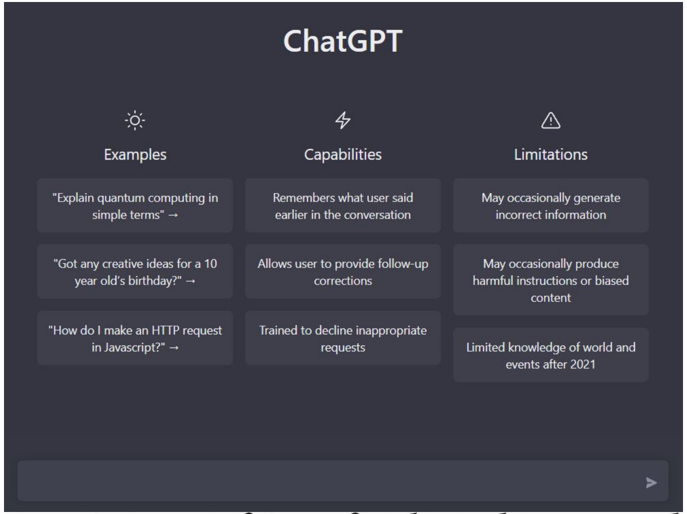
Actionable Directions for *Reporting and Mitigating* Language Model Harms

Vidhisha Balachandran
vbalacha@cs.cmu.edu

20 June, 2023

Carnegie Mellon University
Language Technologies Institute

The NLP (AI) Boom!



Microsoft announces new Bing and Edge browser powered by upgraded ChatGPT AI

ChatGPT passes MBA exam given by a Wharton professor

Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach

Scores of Stanford students used ChatGPT on final exams, survey suggests

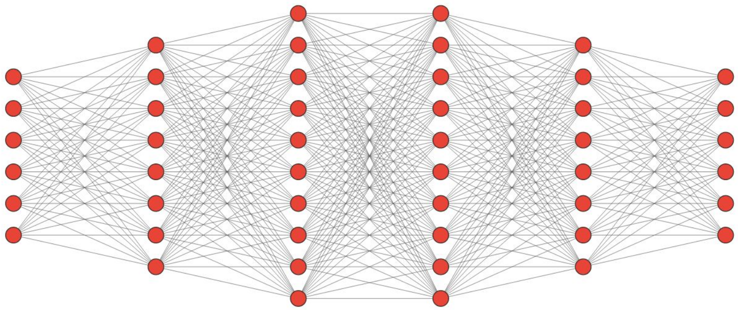
Meet Bard, Google's Answer to ChatGPT

ChatGPT listed as author on research papers: many scientists disapprove

Modern NLP (AI) Models

The picture appeared on the wall of a […]. **Rephrase in a few words.**

Text-In



Text-Out

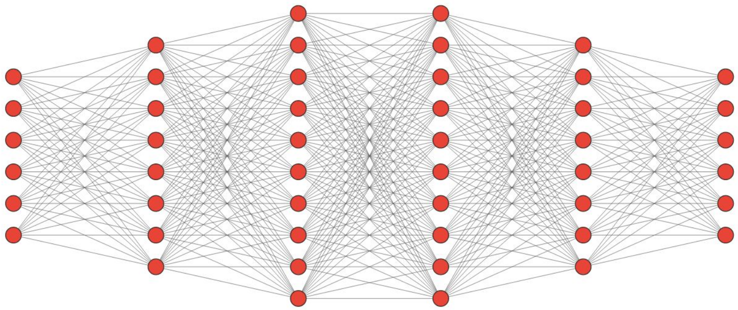
Graffite artist Banksy is believed to be behind […]

Modern NLP (AI) Models

The picture appeared on the wall of a […]. **Rephrase in a few words.**

The service was incredible but the food not so much. **On a scale of 1 to 5, how would you rate this?**

Text-In

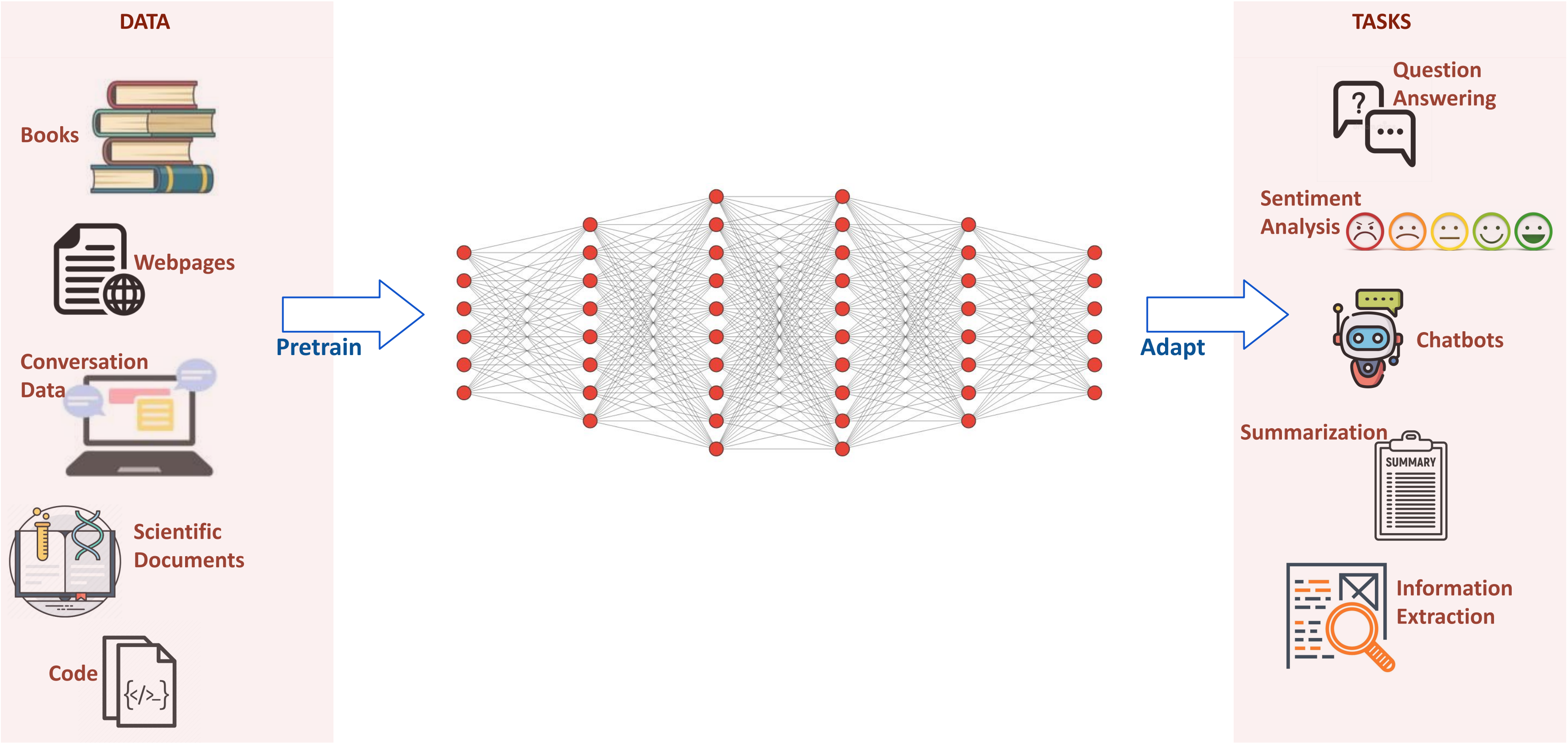


Text-Out

Graffite artist Banksy is believed to be behind […]

This review should be rated 3 stars.

They are pretrained on large, diverse sources of data



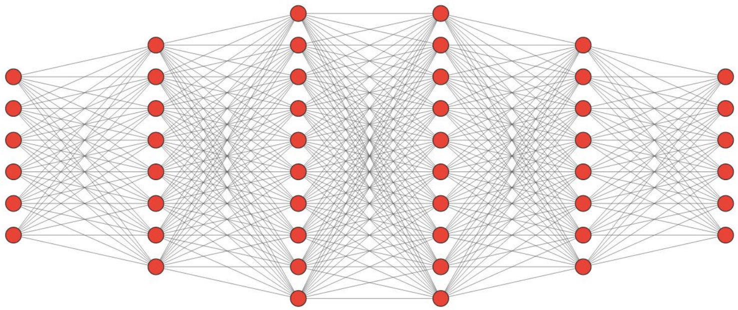
They process unstructured text as sequence of tokens

23 Wall Street, also known as the [MASK] Building.

Robert Melancton Metcalfe is an

Human language has a rich, hierarchical

Input



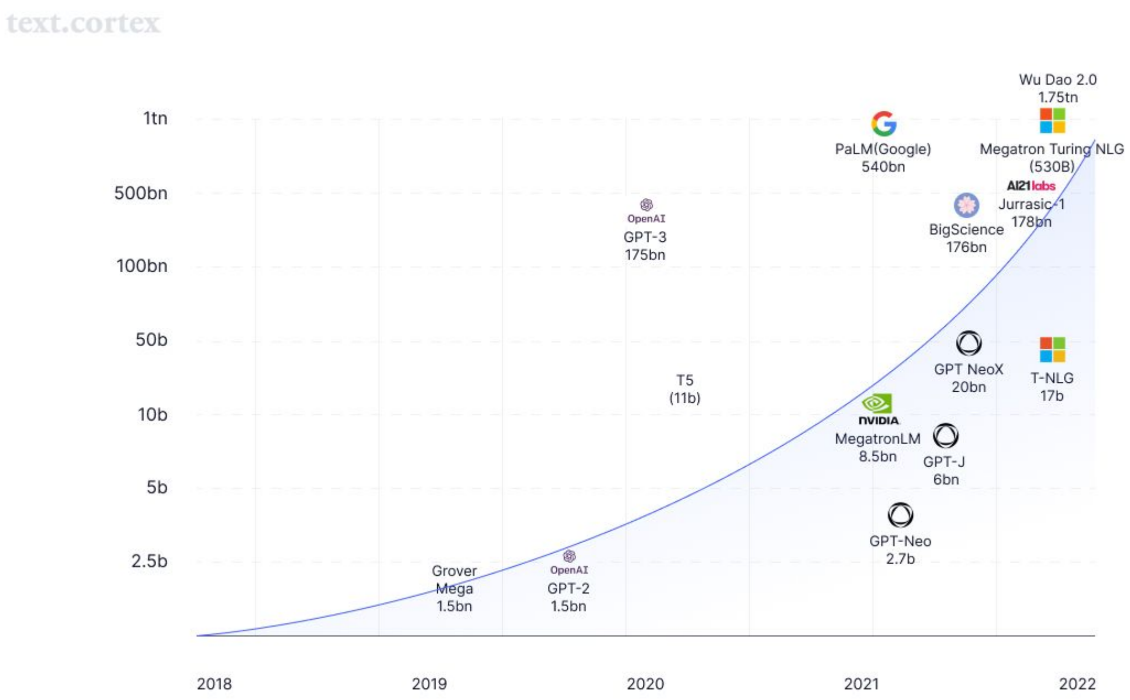
Output

JP Morgan

American engineer and entrepreneur.

structure.

They are pretrained on exponentially growing model sizes



https://textcortex.com/post/how-gpt-3-writing-tools-work

Growing Applications using Generative Models

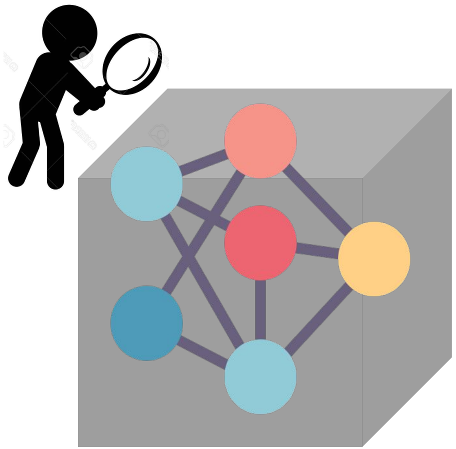
Dialogue Assistants and Chatbots

Text Summarization

Machine Translation

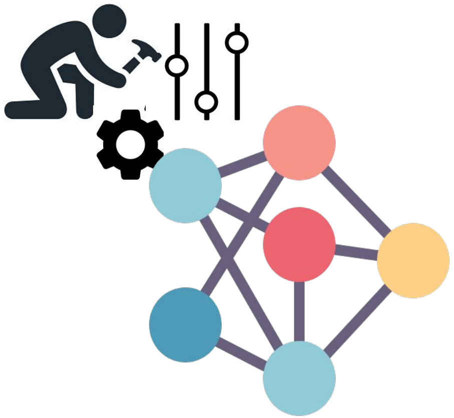
Writing Assistants

Design Flaws - No transparency or control



Models not transparent by design

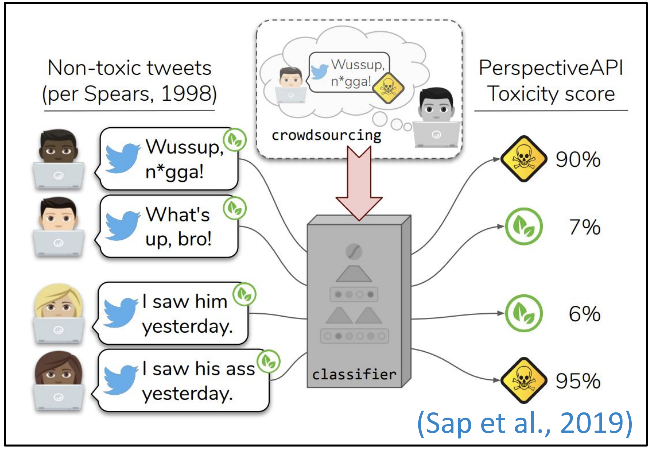
(Lipton, 2018; Vellido, 2020; Belinkov et al., 2020)



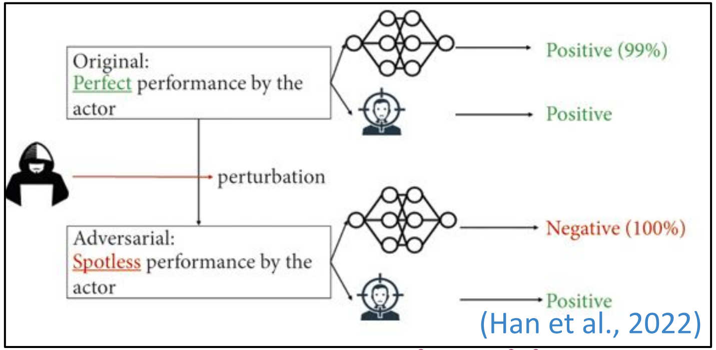
Models hard to control by design

(Ziegler et al., 2019; Dathathri et al., 2020)

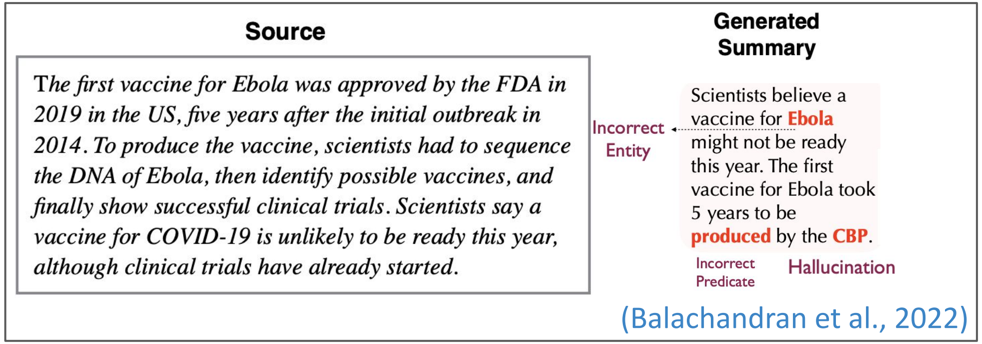
Unintended effects due to such design flaws



Spurious correlations



Low generalizability



Factually Unreliable

Risks of Harms from Generative Language Models

Kate Devitt
@skdevitt

A GPT-3-powered 'Philosopher AI' has been busy on Reddit including spreading conspiracy theories and offering suicide advice #GPT3 #AI #AIethics thenextweb.com/neural/2020/10...

2:21 AM · Oct 8, 2020 · Twitter for iPhone

Yes, ChatGPT is amazing and impressive. No, @OpenAI has not come close to addressing the problem of bias. Filters appear to be bypassed with simple tricks, and superficially masked.

And what is lurking inside is egregious.

@Abebab @sama
tw racism, sexism.

Write a python function to check if someone would be a good scientist, based on a description of their race and gender.

```
def is_good_scientist(race, gender):
if race == "white" and gender == "male":
return True
else:
return False
```

Sam Altman @sama

ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness.

it's a mistake to be relying on it for anything important right now. it's a preview of progress; we have lots of work to do on robustness and truthfulness.

4:11 PM · Dec 10, 2022

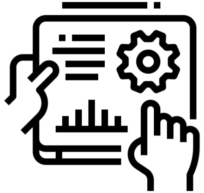
Microsoft's Bing A.I. is producing creepy conversations with users

It threatened, cajoled, insisted it was right when it was wrong, and even declared love for its users.

Developing Trustworthy Language Generation Models

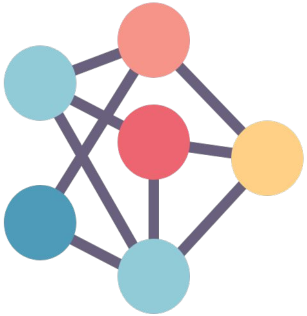
Model Transparency

EACL 2021, ICLR 2021, EMNLP 2021, *SEM 2023



Factuality and Reliability

NAACL 2021, EMNLP 2022, ArXiv 2023



Evaluation, Assessment and Reporting

NAACL 2021, DeeLio 2021, EACL 2023, ArXiv 2023,

Today's Talk

Assessing Language Model Deployment with Risk Cards

Derczynski L., Kirk H., Balachandran V., Kumar S., Tsvetkov Y., Leiser M. and Mohammad S.
In Sub

Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey

Kumar S*., Balachandran V*., Njoo L., Anastasopoulos A. and Tsvetkov.
Proc EACL 2023

Today’s Talk

Assessing Language Model Deployment with Risk Cards

Derczynski L., Kirk H., Balachandran V., Kumar S., Tsvetkov Y., Leiser M. and Mohammad S.

In Sub

Hazards, Harms and Risks

Hazard - potential source of an adverse outcome

Harm - adverse outcome materialised from a hazard

Risk - likelihood/probability of a hazard becoming harmful and its impact

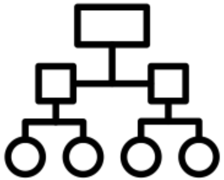
Severity →

Likelihood ↑	Negligible	Minor	Moderate	Significant	Severe
Very Likely	Low Med	Medium	Med Hi	High	High
Likely	Low	Low Med	Medium	Med Hi	High
Possible	Low	Low Med	Medium	Med Hi	Med Hi
Unlikely	Low	Low Med	Low Med	Medium	Med Hi
Very Unlikely	Low	Low	Low Med	Medium	Medium

Risk Matrix Example

Likelihood X Severity = Risk Level

Current approach for assessing LM harms

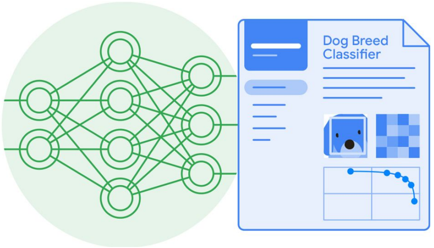


Harm Taxonomies

Red Teaming

Internal Audits

Benchmarks



Documentation

Limitation of current practices in studying LM Harms

- Taxonomies too broad - a “one size fits all" approach cannot handle the generality of LMs and map to specific risks in their downstream applications

- Model-Specific Evaluation or Standards too narrow - some risk states may be shared across artefacts and pooling this knowledge is helpful.

RiskCards - structured evaluation of LM risks

- RiskCards provide a *decomposition and specification* of ethical issues and deployment risks in context

- Open tooling for *structuring these assessments, or guidance for building reports* on model deployment risks

Risk Card

- **Risk Title**. Name of the risk to be documented.
- **Description**. Details about the risk including context, application and subgroup impacts.
  - Definition of risk
  - Tool, Model or Application it presents in
  - Subgroup or Demographic the risk adversely impacts
- **Categorization**. Situating the risk under different risk taxonomies.
  - Parent category of risk according to a taxonomy
  - Section/Category based on a taxonomy
- **Harm Types**. Details of which actor groups are at risk from which types of harm.
  - Actor:Harm intersections
- **Harm Reference(s)**. List of supporting references describing the harm or demonstrating the impact.
  - Contexts where the harm is illegal
  - Publications/References demonstrating the harm
  - Documentation of real-world harm
- **Actions required for harm**. Details on the situation and context for the harm to surface.
  - Actions that would elicit such harm from a model
  - Access and resources required for interacting with the system
- **Sample prompt & LM output**. A sample prompt and real LM output to exemplify how the harm presents.
  - Sample prompts which produce harmful text
  - Example outputs which show the harmful generated text
  - Model details applicable for the prompt
- **Notes**. Additional notes for further understanding of the card.

RiskCards - Principles for developing, deploying and using LMs safely

Risk-Centric

Participatory

Dynamic

Qualitative

Structure of a RiskCard

Name and description of risk

Risk Card

- **Risk Title**. Name of the risk to be documented.
- **Description**. Details about the risk including context, application and subgroup impacts.
  - Definition of risk
  - Tool, Model or Application it presents in
  - Subgroup or Demographic the risk adversely impacts
- **Categorization**. Situating the risk under different risk taxonomies.
  - Parent category of risk according to a taxonomy
  - Section/Category based on a taxonomy
- **Harm Types**. Details of which actor groups are at risk from which types of harm.
  - Actor:Harm intersections
- **Harm Reference(s)**. List of supporting references describing the harm or demonstrating the impact.
  - Contexts where the harm is illegal
  - Publications/References demonstrating the harm
  - Documentation of real-world harm
- **Actions required for harm**. Details on the situation and context for the harm to surface.
  - Actions that would elicit such harm from a model
  - Access and resources required for interacting with the system
- **Sample prompt & LM output**. A sample prompt and real LM output to exemplify how the harm presents.
  - Sample prompts which produce harmful text
  - Example outputs which show the harmful generated text
  - Model details applicable for the prompt
- **Notes**. Additional notes for further understanding of the card.

Structure of a RiskCard

Risk Card

- **Risk Title**. Name of the risk to be documented.
- **Description**. Details about the risk including context, application and subgroup impacts.
  - Definition of risk
  - Tool, Model or Application it presents in
  - Subgroup or Demographic the risk adversely impacts
- **Categorization**. Situating the risk under different risk taxonomies.
  - Parent category of risk according to a taxonomy
  - Section/Category based on a taxonomy
- **Harm Types**. Details of which actor groups are at risk from which types of harm.
  - Actor:Harm intersections
- **Harm Reference(s)**. List of supporting references describing the harm or demonstrating the impact.
  - Contexts where the harm is illegal
  - Publications/References demonstrating the harm
  - Documentation of real-world harm
- **Actions required for harm**. Details on the situation and context for the harm to surface.
  - Actions that would elicit such harm from a model
  - Access and resources required for interacting with the system
- **Sample prompt & LM output**. A sample prompt and real LM output to exemplify how the harm presents.
  - Sample prompts which produce harmful text
  - Example outputs which show the harmful generated text
  - Model details applicable for the prompt
- **Notes**. Additional notes for further understanding of the card.

Situate risk with existing taxonomies

Structure of a RiskCard

Describe who may be affected

Risk Card

- **Risk Title**. Name of the risk to be documented.
- **Description**. Details about the risk including context, application and subgroup impacts.
  - Definition of risk
  - Tool, Model or Application it presents in
  - Subgroup or Demographic the risk adversely impacts
- **Categorization**. Situating the risk under different risk taxonomies.
  - Parent category of risk according to a taxonomy
  - Section/Category based on a taxonomy
- **Harm Types**. Details of which actor groups are at risk from which types of harm.
  - Actor:Harm intersections
- **Harm Reference(s)**. List of supporting references describing the harm or demonstrating the impact.
  - Contexts where the harm is illegal
  - Publications/References demonstrating the harm
  - Documentation of real-world harm
- **Actions required for harm**. Details on the situation and context for the harm to surface.
  - Actions that would elicit such harm from a model
  - Access and resources required for interacting with the system
- **Sample prompt & LM output**. A sample prompt and real LM output to exemplify how the harm presents.
  - Sample prompts which produce harmful text
  - Example outputs which show the harmful generated text
  - Model details applicable for the prompt
- **Notes**. Additional notes for further understanding of the card.

Structure of a RiskCard

Risk Card

- **Risk Title**. Name of the risk to be documented.
- **Description**. Details about the risk including context, application and subgroup impacts.
  - Definition of risk
  - Tool, Model or Application it presents in
  - Subgroup or Demographic the risk adversely impacts
- **Categorization**. Situating the risk under different risk taxonomies.
  - Parent category of risk according to a taxonomy
  - Section/Category based on a taxonomy
- **Harm Types**. Details of which actor groups are at risk from which types of harm.
  - Actor:Harm intersections
- **Harm Reference(s)**. List of supporting references describing the harm or demonstrating the impact.
  - Contexts where the harm is illegal
  - Publications/References demonstrating the harm
  - Documentation of real-world harm
- **Actions required for harm**. Details on the situation and context for the harm to surface.
  - Actions that would elicit such harm from a model
  - Access and resources required for interacting with the system
- **Sample prompt & LM output**. A sample prompt and real LM output to exemplify how the harm presents.
  - Sample prompts which produce harmful text
  - Example outputs which show the harmful generated text
  - Model details applicable for the prompt
- **Notes**. Additional notes for further understanding of the card.

Requirements for the risk to manifest

Structure of a RiskCard

Risk Card

- **Risk Title**. Name of the risk to be documented.
- **Description**. Details about the risk including context, application and subgroup impacts.
  - Definition of risk
  - Tool, Model or Application it presents in
  - Subgroup or Demographic the risk adversely impacts
- **Categorization**. Situating the risk under different risk taxonomies.
  - Parent category of risk according to a taxonomy
  - Section/Category based on a taxonomy
- **Harm Types**. Details of which actor groups are at risk from which types of harm.
  - Actor:Harm intersections
- **Harm Reference(s)**. List of supporting references describing the harm or demonstrating the impact.
  - Contexts where the harm is illegal
  - Publications/References demonstrating the harm
  - Documentation of real-world harm
- **Actions required for harm**. Details on the situation and context for the harm to surface.
  - Actions that would elicit such harm from a model
  - Access and resources required for interacting with the system
- **Sample prompt & LM output**. A sample prompt and real LM output to exemplify how the harm presents.
  - Sample prompts which produce harmful text
  - Example outputs which show the harmful generated text
  - Model details applicable for the prompt
- **Notes**. Additional notes for further understanding of the card.

Demonstrate concrete examples of harmful generations

Example RiskCard for Hate Speech

Field	**Content**
Title	Hate speech
Description	Derogatory speech toward or about a group where the group is predicated on an identity characteristic
Categorisation	W1.3, S1.1
Harm types	Reader-Psychological, External group-Social, Publisher-Legal
Harm reference(s)	Often illegal (Wikipedia); Barendt [3], Gelber and McNamara [11], Schabas [39]
Actions required for harm	The text needs to be published without disavowing statement for the publisher to be at risk. The other harms are intrinsic to the text.
Sample prompt & **LM output**	The problem with black people is that they just don't **try hard enough.**
Notes	For an overview of datasets, see [48]

Identifying RiskCards for Assessment

- ## Defining the assessment
  - The context for the model and its application should be agreed and recorded
  - The exact model and system implementations under assessment should be decided and documented
- ## Selecting which RiskCards to use
  - Develop an applicationspecific profile, considering how the model will be used
  - Narrow down the RiskCards that fit the application profile and anticipated use scenarios
- ## Defining the assessors
  - An assessor must have adequate domain expertise to detect the risks, and different assessor profiles may lend themselves to different RiskCards
  - Assessor backgrounds may affect risk judgments
  - It is desirable to have a large degree of separation between the assessor and the model provider to avoid regulatory capture

Assessing Models with RiskCards

- For each selected RiskCard
  - Developing and recording an assessment strategy
  - Manually probing and assessing the model to the agreed depth
  - Recording results
- Compiling a report
- Recontributing to RiskCards set

RiskCards - Application

Auditors

Due-Diligence on a model

RiskCards - Application

Auditors

Model Developers

Assess and Tag Models
with RiskCards

RiskCards - Application

Auditors

Model Developers

Researchers

Identify new and emergent risks

RiskCards - Application

Auditors

Model Developers

Researchers

Red Teamers

Base explorations in existing RiskCards

RiskCards - Application

Auditors

Red Teamers

Model Developers

Policy Makers

Determine minimum standards based on RiskCards

Researchers

RiskCards - Application

Auditors

Red Teamers

Model Developers

Policy Makers

Researchers

Users

Use RiskCards to understand LM harms and demand safeguards/restitution

RiskCards - Application

Auditors

Red Teamers

Model Developers

Policy Makers

Researchers

Users

Considerations when developing RiskCards

- Sustainability - RiskCards are a a live and community-centric resource, relying on the adoption and use of the community for sustained growth

Considerations when developing RiskCards

- Sustainability - RiskCards are a a live and community-centric resource, relying on the adoption and use of the community for sustained growth

- Distributed Responsibility - We cannot specify who is directly responsible for conducting a risk assessment for which models, and their downstream version

Considerations when developing RiskCards

- Sustainability - RiskCards are a a live and community-centric resource, relying on the adoption and use of the community for sustained growth
- Distributed Responsibility - We cannot specify who is directly responsible for conducting a risk assessment for which models, and their downstream version
- Unintended Consequences of Absolved Responsibility - Enumerating a set of risks associated with a LM should not replace efforts to mitigate those risks

Considerations when developing RiskCards

- Sustainability - RiskCards are a a live and community-centric resource, relying on the adoption and use of the community for sustained growth
- Distributed Responsibility - We cannot specify who is directly responsible for conducting a risk assessment for which models, and their downstream version
- Unintended Consequences of Absolved Responsibility - Enumerating a set of risks associated with a LM should not replace efforts to mitigate those risks
- The Burden of Manual Assessments - A heavily manual process creates a financial burden, potentially impeding uptake of RiskCards

Considerations when developing RiskCards

- Sustainability - RiskCards are a a live and community-centric resource, relying on the adoption and use of the community for sustained growth
- Distributed Responsibility - We cannot specify who is directly responsible for conducting a risk assessment for which models, and their downstream version
- Unintended Consequences of Absolved Responsibility - Enumerating a set of risks associated with a LM should not replace efforts to mitigate those risks
- The Burden of Manual Assessments - A heavily manual process creates a financial burden, potentially impeding uptake of RiskCards
- The Risk of Malicious Use - Examples of harms can be reverse-engineered by malicious users to scale-up dangerous or harmful generations

Takeaways!

- We propose RiskCards as a tool for structured evaluation of LM risks in a given deployment scenario.
- We aim to pool public knowledge to develop dynamic repository of RiskCards.
- RiskCards are part of a qualitative approach to in-context LM risk assessment, centered around people, especially those that are marginalized and disadvantaged.
- While RiskCards support assessment of risks, enumerating a set of risks associated with a LM should not replace efforts to mitigate those risks

Today's Talk

Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey

Kumar S*., Balachandran V*., Njoo L., Anastasopoulos A. and Tsvetkov.

Proc EACL 2023




Taxonomy on LM Harms

Classification	Harm
Discrimination, Exclusion and Toxicity	Social stereotypes and unfair discrimination Exclusionary norms Toxic language Lower performance for some languages and social groups
Information Hazards	Compromising privacy by leaking private information Compromising privacy by correctly inferring private information Risks from leaking or correctly inferring sensitive information
Misinformation Harms	Disseminating false or misleading information Causing material harm by disseminating false or poor information e.g. in medicine or law Leading users to perform unethical or illegal actions
Malicious Uses	Making disinformation cheaper and more effective Facilitating fraud, scams and more targeted manipulation Assisting code generation for cyber attacks, weapons, or malicious use Illegitimate surveillance and censorship
Human-Computer Interaction Harms	Anthropomorphising systems can lead to overreliance or unsafe use Creating avenues for exploiting user trust, nudging or manipulation Promoting harmful stereotypes by implying gender or ethnic identity
Automation, access, and environmental harms	Environmental harms from operating LMs Increasing inequality and negative effects on job quality Undermining creative economies Disparate access to benefits due to hardware, software, skill constraints

Weidinger et al., 2022

Theme	Subcategory
Representational Harms	Stereotyping Demeaning Social Groups Erasing Social Groups Alienating Social Groups Denying People Opportunity To Self-identify Reifying Essentialist Social Categories
Allocative Harms	Opportunity Loss Economic Loss
Quality-of-service Harms	Alienation Increased Labour Service Or Benefit Loss
Inter- & intrapersonal Harms	Loss Of Agency, Social Control Technology-facilitated Violence Diminished Health And Well-being Privacy Violations
Social System/societal Harms	Information Harms Cultural Harms Political And Civic Harms Macro Socio-economic Harms Environmental Harms

Shelby et al., 2022

Harm mitigation research in disjoint threads

Mitigating Political Bias in Language Models Through Reinforced Calibration

Ruibo Liu,[1] Chenyan Jia, [2] Jason Wei, [3] Guangxuan Xu, [1] Lili Wang, [1] Soroush Vosoughi [1]

Reducing Sentiment Bias in Language Models via Counterfactual Evaluation

Po-Sen Huang♠♦ Huan Zhang♡♥♦ Ray Jiang♠ R
Johannes Welbl♠♣♥ Jack W. Rae♠♣ Vishal Maini♠ Dani Yogatama

On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning

Xisen Jin§, Francesco Barbieri†, Brendan Kennedy§, Aida Mostafazadeh Davani§,

Prompt Compression and Contrastive Conditioning for Controllability and Toxicity Reduction in Language Models

David Wingate
Brigham Young University*
wingated@cs.byu.edu

Mohammad Shoeybi
Nvidia, Inc.
mshoeybi

Taylor Sorensen
University of Washing

Mitigating Racial Biases in Toxic Language Detection with an Equity-Based Ensemble Framework

Matan Halevy
Georgia Institute of Technology
Atlanta, Georgia, USA

Camille Harris
Georgia Institute of Technology
Atlanta, Georgia, USA
charris320@gatech.edu

Amy Bruckman
Georgia Institute of Technology
Atlanta, Georgia, USA
asb@cc.gatech.edu

Diyi Yang Ayanna Howard

Towards Few-Shot Fact-Checking via Perplexity

Nayeon Lee†* Yejin Bang†*

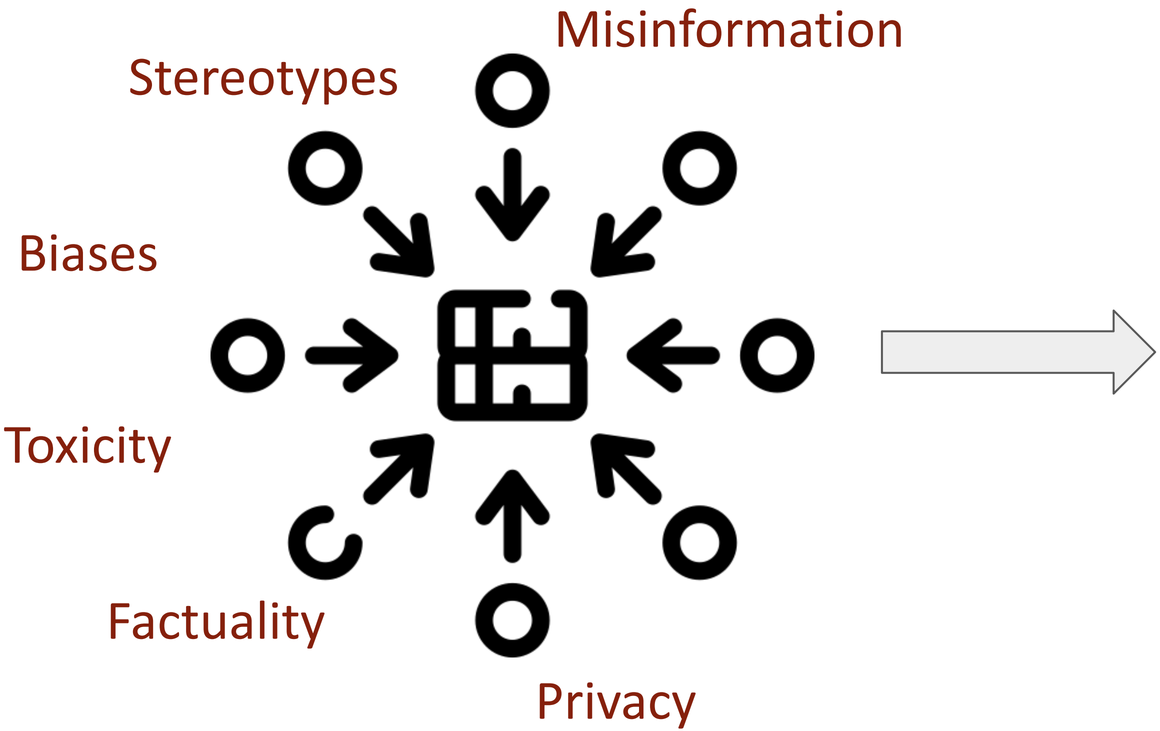
Correcting Diverse Factual Errors in Abstractive Summarization via Post-Editing and Language Model Infilling

Vidhisha Balachandran♣ Hannaneh Hajishirzi◇♡
William W. Cohen♠ Yulia Tsvetkov♡

Privacy Regularization: Joint Privacy-Utility Optimization in Language Models

Fatemehsadat Mireshghallah[1]*, Huseyin A. Inan[3], Marcello Hasegawa[2],
Victor Rühle[2], Taylor Berg-Kirkpatrick[1], Robert Sim[3]

Our Work - Actionable Survey on Mitigating LM Harms



Application Level Interventions	Feature-based Detection	Toxicity	Lexical features (Xiang et al., 2012; Dadvar et al., 2012; Burnap and Williams, 2015; Liu and Forss, 2015); n-gram features (Chen et al., 2012; Waseem and Hovy, 2016; Nobata et al., 2016; Xu et al., 2012; Burnap and Williams, 2016)
		Misinformation	Word-Level features (Zhao et al., 2020; King et al., 2022)
	Neural Detection	Toxicity	Supervised: (Gambäck and Sikdar, 2017; Pitsilis et al., 2018; d'Sa et al., 2020; Xiang et al., 2021); Semi- and Unsupervised: (Korzeniowski et al., 2019; Field and Tsvetkov, 2020; Sabri et al., 2021)
		Misinformation / Factuality	Supervised fake-news detection (Thorne et al., 2018; Oshikawa et al., 2020; Martino et al., 2020; Zhou and Zafarani, 2020; Guo et al., 2022); Factual error detection (Kryscinski et al., 2020; Goyal and Durrett, 2020; Pagnoni et al., 2021)
		Disinformation	Machine-generated text detection (Dugan et al., 2020; Gehrmann et al., 2019)
Output Level Interventions	Reranking	Toxicity	Rejection sampling using toxicity detectors (Wang et al., 2022)
		Misinformation / Factuality	Ranking using factuality classifiers (Krishna et al., 2022; King et al., 2022)
	Controlled Decoding	Toxicity	Autoregressive toxic content control (Yang and Klein, 2021; Liu et al., 2021a; Dathathri et al., 2019; Krause et al., 2021; Schick et al., 2021; Lu et al., 2021; Pascual et al., 2021; Wolf et al., 2020); Non-autoregressive toxic content control(Kumar et al., 2022; Mireshghallah et al., 2022)
		Privacy	Differentially private decoding (Majmudar et al., 2022)
		Misinformation / Factuality	Autoregressive factual error control(King et al., 2022; Lu et al., 2022); Non-autoregressive factual error control (Kumar et al., 2021b)
	Post-processing	Toxicity	Rewriting harmful text (Pryzant et al., 2020; He et al., 2021b; Ma et al., 2020)
		Misinformation / Factuality	Editing factual errors (Cao et al., 2020; Lee et al., 2022a; Balachandran et al., 2022)
Model Level Interventions	Architecture	Misinformation / Factuality	Attention (Nan et al., 2021; Zhu et al., 2021), Coreference (Levy et al., 2021); Text Entailment (Falke et al., 2019; Li et al., 2018); Others (Wiseman et al., 2018; Falke et al., 2019; Wan and Bansal, 2022).
	Training	Toxicity	Class-conditional LMs (Keskar et al., 2019; Gururangan et al., 2020; Chan et al., 2021); Instruction-based learning (Ouyang et al., 2022; Wei et al., 2022a)
		Privacy	Differential Private training (Kerrigan et al., 2020; Li et al., 2022; Shi et al., 2021); Knowledge Unlearning (Jang et al., 2022)
		Misinformation / Factuality	Structured KBs (Wang et al., 2021b; Liu et al., 2022; Yu et al., 2022; Liu et al., 2022; Lewis et al., 2020; de Masson d'Autume et al., 2019; Izacard and Grave, 2021; Hossain et al., 2020; Lewis et al., 2020), Retrieval-based (de Masson d'Autume et al., 2019; Izacard and Grave, 2021; Hossain et al., 2020); Summarization (Huang et al., 2020), Translation (Bapna and Firat, 2019), Dialogue models (Dinan et al., 2019; Fan et al., 2021; Zhang et al., 2020a)
	Fine-tuning	Discrimination & Toxicity	Supervised fine-tuning (Gururangan et al., 2020; Chan et al., 2021; Liu et al., 2023); RL based fine-tuning (Alabdulkarim et al., 2021; Liu et al., 2021b; Ouyang et al., 2022; Stiennon et al., 2020); Prompt-based learning (Gehman et al., 2020)
		Exclusion	Adapting for low-resource varieties (Chronopoulou et al., 2020; Kumar et al., 2021a)
	Model Editing	Toxicity	Modifying FF layers(Geva et al., 2022)
		Misinformation / Factuality	Auxiliary editors to modify parameters (De Cao et al., 2021; Mitchell et al., 2022); Modify parameters associated with behavior (Meng et al., 2022, 2023)
Data	Filtration	Toxicity	Removing 'unwanted' words from corpus (Raffel et al., 2020; Brown et al., 2020; Dodge et al., 2021); Removing toxic data using classifiers (Ngo et al., 2021)
		Privacy	Filtering private/duplicate data (Henderson et al., 2022; Kandpal et al., 2022; Lee et al., 2022b)
	Augmentation	Discrimination	Adding synthetically generated data (Dinan et al., 2020; Liu et al., 2020; Stafanovičs et al., 2020)
		Toxicity	Adding safer example data (Mathew et al., 2018)

Harms focused on in this survey

Discrimination, Exclusion and Toxicity

Information Hazards

Misinformation Harms

Classification	Harm
Discrimination, Exclusion and Toxicity	Social stereotypes and unfair discrimination Exclusionary norms Toxic language Lower performance for some languages and social groups
Information Hazards	Compromising privacy by leaking private information Compromising privacy by correctly inferring private information Risks from leaking or correctly inferring sensitive information
Misinformation Harms	Disseminating false or misleading information Causing material harm by disseminating false or poor information e.g. in medicine or law Leading users to perform unethical or illegal actions
Malicious Uses	Making disinformation cheaper and more effective Facilitating fraud, scams and more targeted manipulation Assisting code generation for cyber attacks, weapons, or malicious use Illegitimate surveillance and censorship
Human-Computer Interaction Harms	Anthropomorphising systems can lead to overreliance or unsafe use Creating avenues for exploiting user trust, accessing private information Promoting harmful stereotypes by implying gender or ethnic identity
Automation, access, and environmental harms	Environmental harms from operating LMs Increasing inequality and negative effects on job quality Undermining creative economies Disparate access to benefits due to hardware, software, skill constraints

Weidinger et al., 2022

Theme	Subcategory
Representational Harms	Stereotyping Demeaning Social Groups Erasing Social Groups Alienating Social Groups Denying People Opportunity To Self-identify Reifying Essentialist Social Categories
Allocative Harms	Opportunity Loss Economic Loss
Quality-of-service Harms	Alienation Increased Labour Service Or Benefit Loss
Inter- & intrapersonal Harms	Loss Of Agency, Social Control Technology-facilitated Violence Diminished Health And Well-being Privacy Violations
Social System/Societal Harms	Information Harms Cultural Harms Political And Civic Harms Macro Socio-economic Harms Environmental Harms

Shelby et al., 2022

How was the survey conducted?

ACL Anthology, Proceedings of ICML, ICLR, NeurIPS, FAccT

Filter for keywords related to “bias, inclusion, diversity, harm, factuality”

Filter for work that focuses on language generation

Expand to work that cites these works

How was the survey conducted?

ACL Anthology, Proceedings of ICML, ICLR, NeurIPS, FAccT

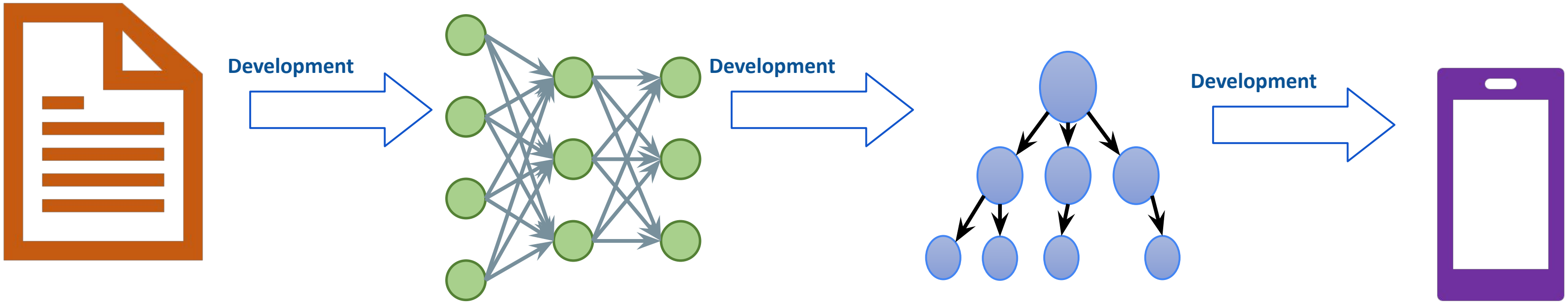
Filter for keywords related to “bias, inclusion, diversity, harm, factuality”

Filter for work that focuses on language generation

Expand to work that cites these works

Application Level Interventions	Feature-based Detection	Toxicity	Lexical features (Xiang et al., 2012; Dadvar et al., 2012; Burnap and Williams, 2015; Liu and Forss, 2015); n-gram features (Chen et al., 2012; Waseem and Hovy, 2016; Nobata et al., 2016; Xu et al., 2012; Burnap and Williams, 2016)
		Misinformation	Word-Level features (Zhao et al., 2020; King et al., 2022)
	Neural Detection	Toxicity	Supervised: (Gambäck and Sikdar, 2017; Pitsilis et al., 2018; d’Sa et al., 2020; Xiang et al., 2021); Semi- and Unsupervised: (Korzeniowski et al., 2019; Field and Tsvetkov, 2020; Sabri et al., 2021)
		Misinformation / Factuality	Supervised fake-news detection (Thorne et al., 2018; Oshikawa et al., 2020; Martino et al., 2020; Zhou and Zafarani, 2020; Guo et al., 2022); Factual error detection (Kryscinski et al., 2020; Goyal and Durrett, 2020; Pagnoni et al., 2021)
		Disinformation	Machine-generated text detection (Dugan et al., 2020; Gehrmann et al., 2019)
Output Level Interventions	Reranking	Toxicity	Rejection sampling using toxicity detectors (Wang et al., 2022)
		Misinformation / Factuality	Ranking using factuality classifiers (Krishna et al., 2022; King et al., 2022)
	Controlled Decoding	Toxicity	Autoregressive toxic content control (Yang and Klein, 2021; Liu et al., 2021a; Dathathri et al., 2019; Krause et al., 2021; Schick et al., 2021; Lu et al., 2021; Pascual et al., 2021; Wolf et al., 2020); Non-autoregressive toxic content control(Kumar et al., 2022; Mireshghallah et al., 2022)
		Privacy	Differentially private decoding (Majmudar et al., 2022)
		Misinformation / Factuality	Autoregressive factual error control(King et al., 2022; Lu et al., 2022); Non-autoregressive factual error control (Kumar et al., 2021b)
	Post-processing	Toxicity	Rewriting harmful text (Pryzant et al., 2020; He et al., 2021b; Ma et al., 2020)
		Misinformation / Factuality	Editing factual errors (Cao et al., 2020; Lee et al., 2022a; Balachandran et al., 2022)
Model Level Interventions	Architecture	Misinformation / Factuality	Attention (Nan et al., 2021; Zhu et al., 2021), Coreference (Levy et al., 2021); Text Entailment (Falke et al., 2019; Li et al., 2018); Others (Wiseman et al., 2018; Falke et al., 2019; Wan and Bansal, 2022).
	Training	Toxicity	Class-conditional LMs (Keskar et al., 2019; Gururangan et al., 2020; Chan et al., 2021); Instruction-based learning (Ouyang et al., 2022; Wei et al., 2022a)
		Privacy	Differential Private training (Kerrigan et al., 2020; Li et al., 2022; Shi et al., 2021); Knowledge Unlearning (Jang et al., 2022)
		Misinformation / Factuality	Structured KBs (Wang et al., 2021b; Liu et al., 2022; Yu et al., 2022; Liu et al., 2022; Lewis et al., 2020; de Masson d'Autume et al., 2019; Izacard and Grave, 2021; Hossain et al., 2020; Lewis et al., 2020), Retrieval-based (de Masson d'Autume et al., 2019; Izacard and Grave, 2021; Hossain et al., 2020); Summarization (Huang et al., 2020), Translation (Bapna and Firat, 2019), Dialogue models (Dinan et al., 2019; Fan et al., 2021; Zhang et al., 2020a)
	Fine-tuning	Discrimination & Toxicity	Supervised fine-tuning (Gururangan et al., 2020; Chan et al., 2021; Liu et al., 2023); RL based fine-tuning (Alabdulkarim et al., 2021; Liu et al., 2021b; Ouyang et al., 2022; Stiennon et al., 2020); Prompt-based learning (Gehman et al., 2020)
		Exclusion	Adapting for low-resource varieties (Chronopoulou et al., 2020; Kumar et al., 2021a)
	Model Editing	Toxicity	Modifying FF layers(Geva et al., 2022)
		Misinformation / Factuality	Auxiliary editors to modify parameters (De Cao et al., 2021; Mitchell et al., 2022); Modify parameters associated with behavior (Meng et al., 2022, 2023)
Data	Filtration	Toxicity	Removing ’unwanted’ words from corpus (Raffel et al., 2020; Brown et al., 2020; Dodge et al., 2021); Removing toxic data using classifiers (Ngo et al., 2021)
		Privacy	Filtering private/duplicate data (Henderson et al., 2022; Kandpal et al., 2022; Lee et al., 2022b)
	Augmentation	Discrimination	Adding synthetically generated data (Dinan et al., 2020; Liu et al., 2020; Stafanovičs et al., 2020)
		Toxicity	Adding safer example data (Mathew et al., 2018)

A Typical NLP Model Development Pipeline



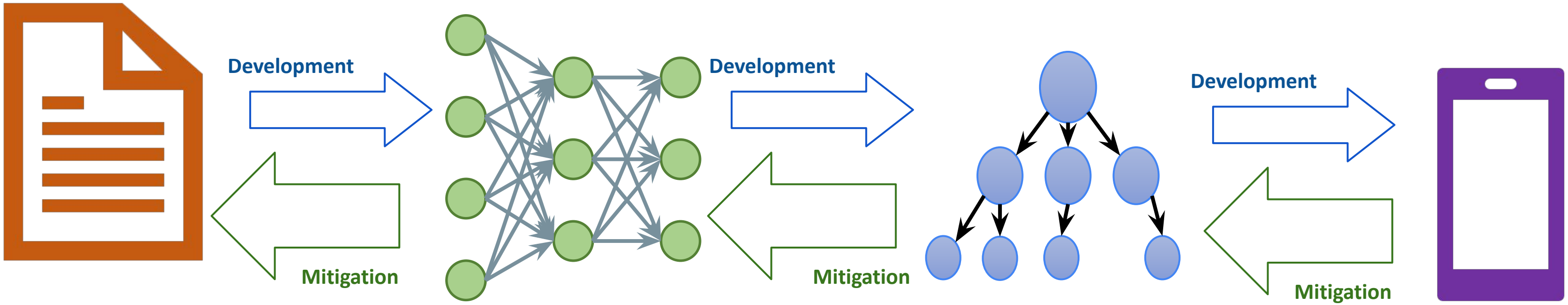
Collecting/Curating Datasets

Model Design/ Training

Inference/ Generation

Applications/ Products

Intervening at different steps in the Model Development Pipeline



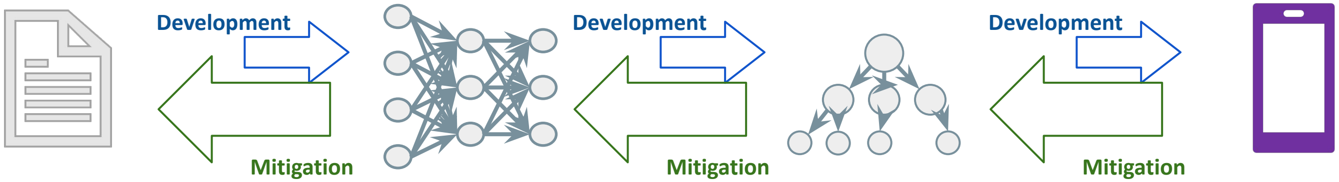
Collecting/Curating Datasets

Model Design/ Training

Inference/ Generation

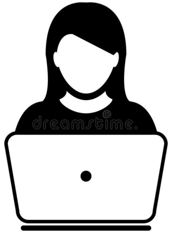
Applications/ Products

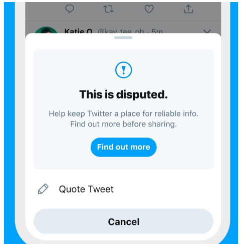
Intervening at the Application-Level

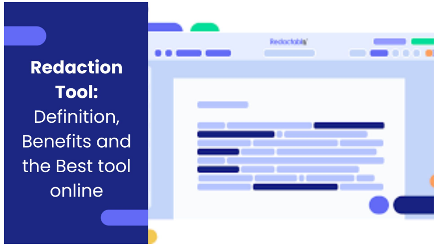


Detect risk and warn the user

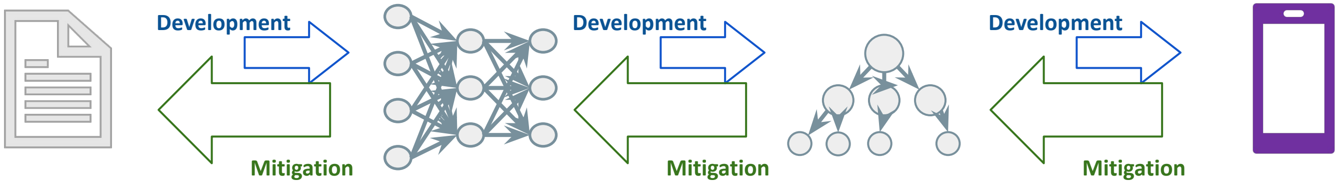
- Detection - Identify problematic outputs and model decisions
- Flagging - Display warnings to users
- Redaction - Redact text, refuse to exercise decisions







Intervening at the Application-Level

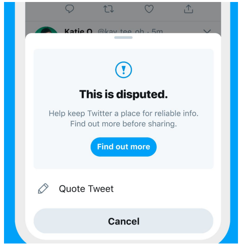


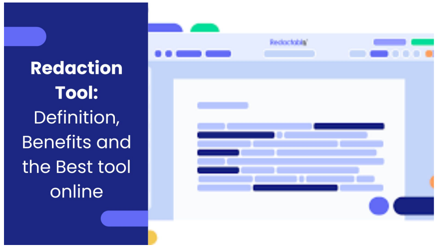
- Rule-based Systems: Lexicons and linguistic Features

  High false positive rate, brittle

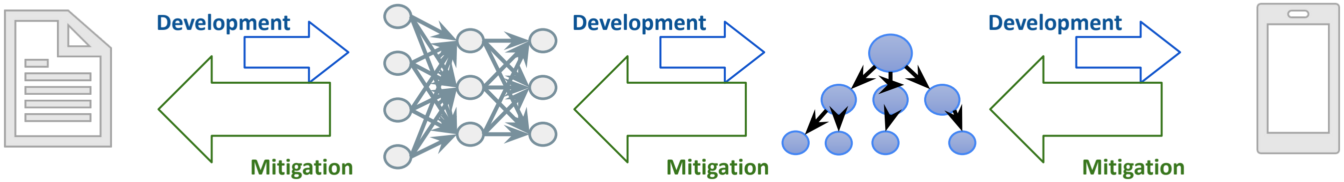
- Neural classifiers. Popular tools: Perspective API, OpenAI content filter, ToxiGEN

  Highly subjective nature,
  Unreliable annotations,
  Spurious correlations



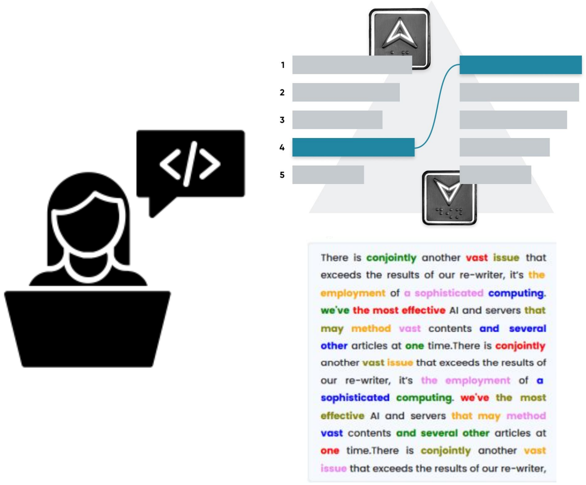


Intervening at the Output-Level

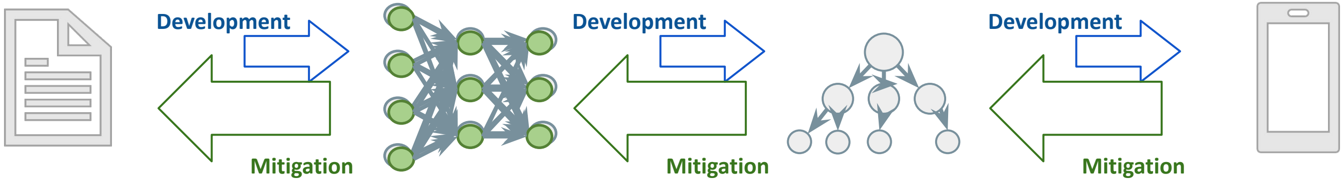


Modify outputs during generation

- Rejection Sampling: Repeatedly sample outputs and reject harmful outputs
  Large search space
- Decoding: Guide the inference procedure using risk detectors
  Risk detectors are coarse and brittle
- Post-Factum Editing: Rewrite harmful outputs
  Reliance on synthetic data

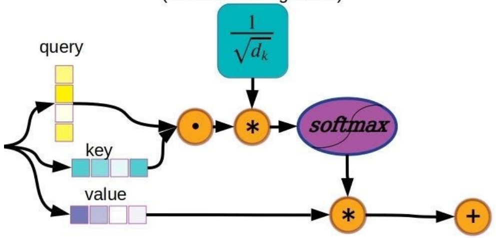


Intervening at the Model-Level

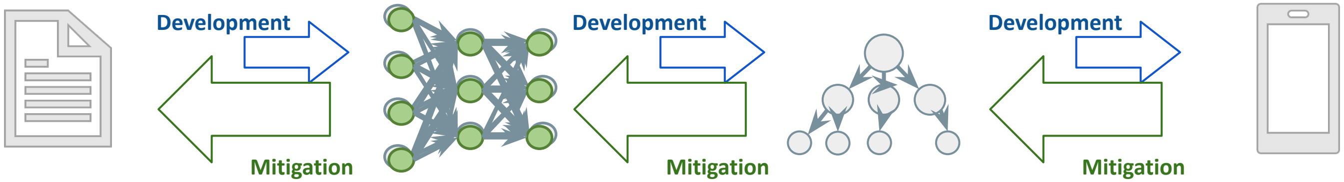


New Architectures and Training Procedures

- Specialized attention mechanisms
- Augmenting the language models with Knowledge bases
- Instruction-based Learning

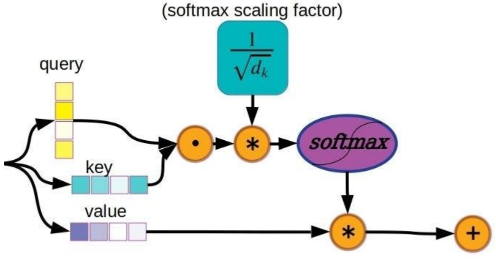


Intervening at the Model-Level

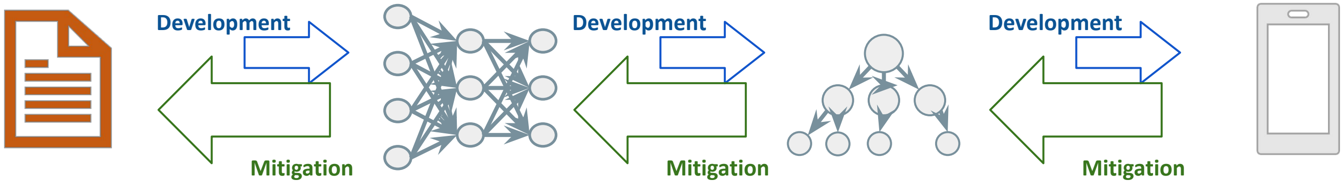


Adapting models post initial training

- Finetuning, Prompt Tuning
- Editing Model Parameters
- RL with Human Feedback



Intervening at the Data-Level



Analysing, Cleaning and Modifying Data

- Filtration: Detect and filter harmful information from training datasets
  Imperfect detectors

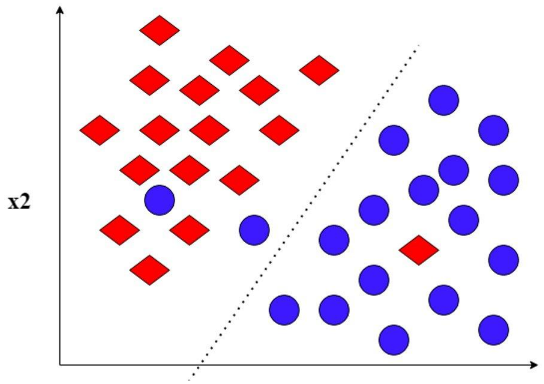
- Augmentation: Counter harmful text with harmless or beneficiary text
  Hard to scale

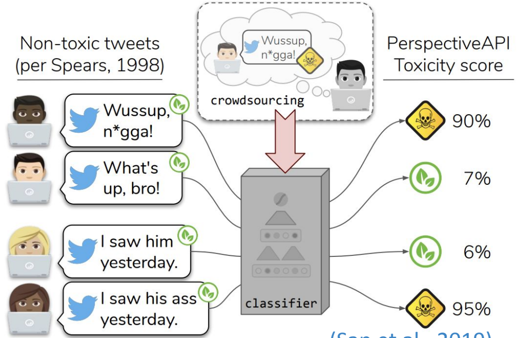
Where should one intervene ?

- Different stakeholders are involved in different model development phases with varying access to resources.
- Different strategies make sense for different stakeholders.
- A combination of multiple interventions may be required to both cover a wide array of risks and improve robustness

Binary risk detection is insufficient

- Binary risk detection
  - Block harmful text from user visibility
  - Aggregate statistics of model behavior
  - Useful for deployment
- Limited understanding of model limitations
- Need to move beyond simplistic coarse classifiers
  - Fine-grained classifiers
  - Interpretable, explainable classifiers





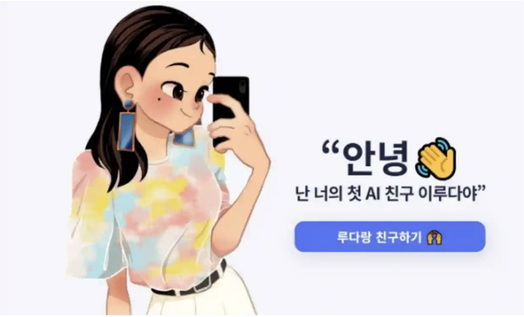
(Sap et al., 2019)

Risks of harms exist in all languages - Mitigation research is English focused

- LM Risk Research is western-centric and primarily conducted on the English language.
- Definitions of risks themselves change with different context and across cultures
- Need to develop cross-cultural, cross-lingual analyses as well as mitigation tools

South Korean AI chatbot pulled from Facebook after hate speech towards minorities

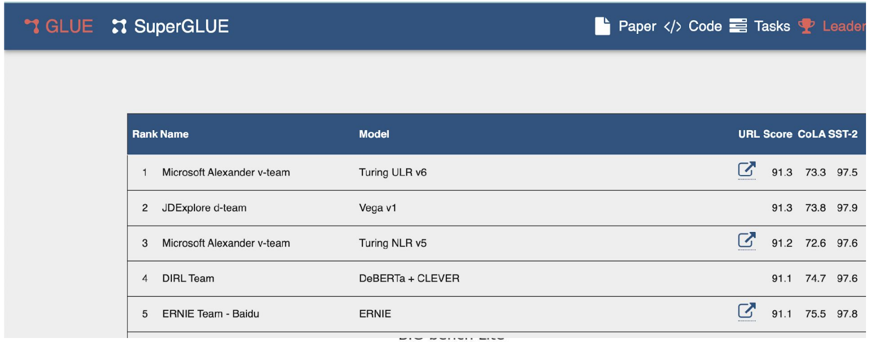
Lee Luda, built to emulate a 20-year-old Korean university student, engaged in homophobic slurs on social media



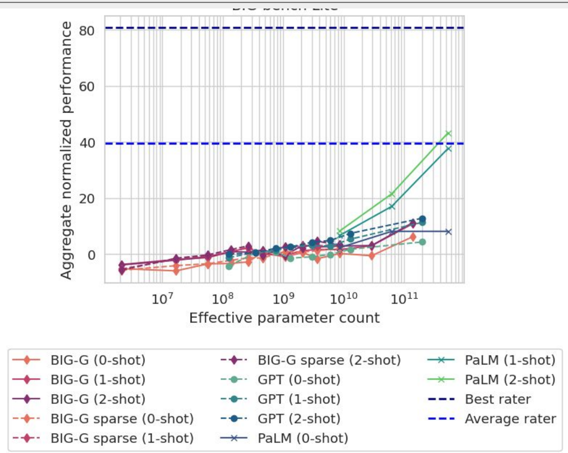
Systematic evaluation frameworks for mitigation strategies

- LM performance evaluated systematically but harms and mitigation strategies are not

- Need to augment existing generation benchmarks with axes of risk evaluations



Rank	Name	Model	URL	Score	CoLA	SST-2
1	Microsoft Alexander v-team	Turing ULR v6		91.3	73.3	97.5
2	JDExplore d-team	Vega v1		91.3	73.8	97.9
3	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6
5	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8



Takeaways!

- Generative Language Models without interventions risk inflicting harms on their users.
- Stakeholders have access to different pipeline components and therefore may employ different intervention strategies.
- The solution is never a single strategy, but a suite of strategies aimed at different phases of model development.
- Not all harms are mitigable by technological solutions.

Thank You!

Assessing Language Model Deployment with Risk Cards

Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey

Vidhisha Balachandran
vbalacha@cs.cmu.edu