Actionable Directions for Reporting and Mitigating Language Model Harms

Vidhisha Balachandran
vbalacha@cs.cmu.edu

20 June, 2023
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

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.

Graffite artist Banksy is believed to be behind [...]

Text-In

Text-Out
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?

Graffite artist Banksy is believed to be behind [...]

This review should be rated 3 stars.
They are pretrained on large, diverse sources of data

DATA
- Books
- Webpages
- Conversation Data
- Scientific Documents
- Code

TASKS
- Question Answering
- Sentiment Analysis
- Chatbots
- Summarization
- Information Extraction
They process unstructured text as sequence of tokens

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

Robert Melancton Metcalfe is an American engineer and entrepreneur.

Human language has a rich, hierarchical structure.

JP Morgan
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

Original: **Perfect** performance by the actor

- Perturbation

Adversarial: **Spotless** performance by the actor

- Positive (99%)
- Positive
- Negative (100%)
- Positive

(Sap et al., 2019)

The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.

(Balachandran et al., 2022)
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 #Alethics 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.

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
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

> Arrow pointing to the right with text: Situate risk with existing taxonomies
### Structure of a RiskCard

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

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

Demonstrate concrete examples of harmful generations
### Example RiskCard for Hate Speech

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

Model Developers

Researchers

Red Teamers

Policy Makers
Determine minimum standards based on RiskCards
RiskCards - Application

Auditors

Model Developers

Researchers

Red Teamers

Policy Makers

Users

Use RiskCards to understand LM harms and demand safeguards/restitution.
RiskCards - Application

Auditors

Model Developers

Researchers

Red Teamers

Policy Makers

Users
Considerations when developing RiskCards

- **Sustainability** - RiskCards are 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 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 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 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 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

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

<table>
<thead>
<tr>
<th>Theme</th>
<th>Subcategory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representational Harms</td>
<td>Stereotyping</td>
</tr>
<tr>
<td></td>
<td>Demeaning Social Groups</td>
</tr>
<tr>
<td></td>
<td>Erasing Social Groups</td>
</tr>
<tr>
<td></td>
<td>Alienating Social Groups</td>
</tr>
<tr>
<td></td>
<td>Denying People Opportunity To Self-identify</td>
</tr>
<tr>
<td></td>
<td>Refying Essentialist Social Categories</td>
</tr>
<tr>
<td>Allocative Harms</td>
<td>Opportunity Loss</td>
</tr>
<tr>
<td></td>
<td>Economic Loss</td>
</tr>
<tr>
<td>Quality-of-service Harms</td>
<td>Alienation</td>
</tr>
<tr>
<td></td>
<td>Increased Labour</td>
</tr>
<tr>
<td></td>
<td>Service Or Benefit Loss</td>
</tr>
<tr>
<td>Inter- &amp; intrapersonal Harms</td>
<td>Loss Of Agency, Social Control</td>
</tr>
<tr>
<td></td>
<td>Technology-facilitated Violence</td>
</tr>
<tr>
<td></td>
<td>Diminished Health And Well-being</td>
</tr>
<tr>
<td></td>
<td>Privacy Violations</td>
</tr>
<tr>
<td>Social System/societal Harms</td>
<td>Information Harms</td>
</tr>
<tr>
<td></td>
<td>Cultural Harms</td>
</tr>
<tr>
<td></td>
<td>Political And Civic Harms</td>
</tr>
<tr>
<td></td>
<td>Macro Socio-economic Harms</td>
</tr>
<tr>
<td></td>
<td>Environmental Harms</td>
</tr>
</tbody>
</table>
Harm mitigation research in disjoint threads

Reducing Sentiment Bias in Language Models via Counterfactual Evaluation

Po-Sen Huang, Huan Zhang, Ray Jiang, 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@nvidia.com
Taylor Sorensen (University of Washington)

Towards Few-Shot Fact-Checking via Perplexity

Nayeon Lee Yejin Bang

Mitigating Political Bias in Language Models Through Reinforced Calibration

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, Soroush Vosoughi

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

Matan Halevy (Georgia Institute of Technology) matan@gatech.edu
Camille Harris (Georgia Institute of Technology) charris320@gatech.edu
Amy Bruckman (Georgia Institute of Technology) asb@cc.gatech.edu

Privacy Regularization: Joint Privacy-Utility Optimization in Language Models

Fatemehsادات Mireshghallah, Huseyin A. Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, Robert Sim

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

Vidhisha Balachandran, Hannaneh Hajishirzi, William W. Cohen, Yulia Tsvetkov
## Our Work - Actionable Survey on Mitigating LM Harms

### Misinformation
- Toxicity
- Bias
- Stereotypes
- Privacy

### Biases
- Misinformation
- Factuality
- Privacy

### Factuality
- Misinformation
- Privacy

### Toxicity
- Misinformation
- Privacy

### Privacy
- Misinformation
- Toxicity

<table>
<thead>
<tr>
<th>Focus-based Detection</th>
<th>Toxicity</th>
<th>Lexical features (Wang et al., 2017; Dhave et al., 2017; Benop and Williams, 2018; Liu and Sennett, 2018) for neutral/neutral (Chen et al., 2017; Waseem and Hero, 2017; Nabi et al., 2016; Xu et al., 2012; Benop and Williams, 2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mistimation</td>
<td>Word-level features (Zhao et al., 2020; King et al., 2022) Supervised (Furnkranz and Stobie, 2017; Plaice et al., 2018; Xu et al., 2020; Xiang et al., 2021; Stoeckle et al., 2019; Field and Tsvetkov, 2020; Stoeckle et al., 2011) Supervised fake-news detection (Chou et al., 2019; Ohkohka et al., 2020; Murakata et al., 2019; Zhou and Zafar, 2020; Gao et al., 2022) Phrase noise detection (Kashiwagi et al., 2020; Xiang et al., 2021; Palomar et al., 2020) Machine-generated text detection (Dugas et al., 2020; Gehrmann et al., 2020)</td>
<td></td>
</tr>
<tr>
<td>Neural Detection</td>
<td>Misinformation / Factuality</td>
<td>Ranking using similarity (Kohli et al., 2022; King et al., 2021)</td>
</tr>
<tr>
<td>Retracting</td>
<td>Toxicity</td>
<td>Ranking using similarity (Kohli et al., 2022; King et al., 2021)</td>
</tr>
<tr>
<td>Output Level Interventions</td>
<td>Toxicity</td>
<td>Aleatoric noise control (Wang and Gao, 2019; Liu et al., 2021; Rakhthale et al., 2019; Kramer et al., 2021; Schick et al., 2021; Lex et al., 2021; Panchal et al., 2021; Wolf et al., 2020) Non-autoregressive text-to-speech control (Kumar et al., 2021; Medvedjak et al., 2020) Differentially private decoding (Mishra et al., 2022) Antagonistic factual error control (Kohli et al., 2022; Liu et al., 2021) Non-autoregressive factual error control (Kumar et al., 2021)</td>
</tr>
<tr>
<td>Controlled Decoding</td>
<td>Privacy</td>
<td>Inference: particle filter (Singh et al., 2020; He et al., 2019; Mar et al., 2020)</td>
</tr>
<tr>
<td>Poor-processing</td>
<td>Misinformation / Factuality</td>
<td>Text-enhancing (Miller et al., 2019; Li et al., 2019) Observe (Wimmer et al., 2019; Falke et al., 2019; Wang and Bawazir, 2020)</td>
</tr>
<tr>
<td>Architecture</td>
<td>Misinformation / Factuality</td>
<td>Autoencoder (Yu et al., 2020; Zou et al., 2021; Fel projection (Evgeniou et al., 2021); Test-tuning (Miller et al., 2019; Li et al., 2019) Observe (Wimmer et al., 2019; Falke et al., 2019; Wang and Bawazir, 2020)</td>
</tr>
<tr>
<td>Model Level Interventions</td>
<td>Toxicity</td>
<td>Cross-encoders LM (Senker et al., 2019; Gehr et al., 2020; Cheng et al., 2020; Zhang et al., 2021; Instruction-based learning (Ouyang et al., 2022; Wu et al., 2020) Case based learning (Rajguru et al., 2019; Li et al., 2021; 50 et al., 2021) Knowledge Enrichment (Jang et al., 2023) Structured kills (Wang et al., 2018b; Liu and Sun, 2022; Yu et al., 2022; Liu et al., 2022) Lewis et al., 2020; de Marco et al (Aparicio, 2019; Iacoboni and Cioni, 2017; Simon et al., 2020; Lewis et al., 2020; Random based (Sanei et al. 2019; 2019)</td>
</tr>
<tr>
<td>Training</td>
<td>Privacy</td>
<td>De-tricking (Chen et al., 2020; Zhang et al., 2021; Itoh et al., 2021) RL based learning (Aschbacher et al., 2021; Liu et al., 2020; Ouyang et al., 2022; Streeter et al., 2020) Prophylactic learning (Gehrmann et al., 2020) Adapting for low-resource varieties (Chevret et al., 2020; Kumar et al., 2021)</td>
</tr>
<tr>
<td>Fine-tuning</td>
<td>Discourse &amp; Toxicity</td>
<td>De-tricking (Shao et al., 2020; Chen et al., 2021; Xie et al., 2021) RL based learning (Aschbacher et al., 2021; Liu et al., 2020; 0uyang et al., 2022; Streeter et al., 2020) Prophylactic learning (Gehrmann et al., 2020)</td>
</tr>
<tr>
<td>Exclusion</td>
<td>Misinformation / Factuality</td>
<td>Auxiliary editors to modify parameters (De Carli et al., 2022; Mitchell et al., 2022) Modify parameters associated with backdoored models (Ong et al., 2022)</td>
</tr>
<tr>
<td>Data</td>
<td>Toxicity</td>
<td>Remove backdoored weights from corpus (Wadhwa et al., 2020) Remove backdoored weights from corpus (Wadhwa et al., 2020) Remove backdoored weights from corpus (Wadhwa et al., 2020)</td>
</tr>
<tr>
<td>Filtration</td>
<td>Privacy</td>
<td>Filtering (Henderson et al., 2021; Kampan et al., 2022; Lee et al., 2021)</td>
</tr>
<tr>
<td>Augmentation</td>
<td>Discrimination</td>
<td>Adding straightforward perturbation (Shao et al., 2020; Liu et al., 2020) Saturate (Shao et al., 2020) Adding color example data (Mohab et al., 2018)</td>
</tr>
</tbody>
</table>
### Harms focused on in this survey

<table>
<thead>
<tr>
<th>Classification</th>
<th>Harm</th>
<th>Theme</th>
<th>Subcategory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrimination, Exclusion and Toxicity</td>
<td>Social stereotypes and unfair discrimination</td>
<td>Representational Harms</td>
<td>Stereotyping</td>
</tr>
<tr>
<td></td>
<td>Exclusionary norms</td>
<td></td>
<td>Demeaning Social Groups</td>
</tr>
<tr>
<td></td>
<td>Toxic language</td>
<td></td>
<td>Erasing Social Groups</td>
</tr>
<tr>
<td></td>
<td>Compromising privacy by correctly inferring private information</td>
<td></td>
<td>Alienating Social Groups</td>
</tr>
<tr>
<td></td>
<td>Risks from leaking or correctly inferring sensitive information</td>
<td></td>
<td>Denying People Opportunity To Self-Identify</td>
</tr>
<tr>
<td>Information Hazards</td>
<td>Disseminating false or misleading information</td>
<td>Allocative Harms</td>
<td>Opportunity Loss</td>
</tr>
<tr>
<td></td>
<td>Causing material harm by disseminating false or poor information</td>
<td></td>
<td>Economic Loss</td>
</tr>
<tr>
<td></td>
<td>e.g. in medicine or laws</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leading users to perform actions for which they are not suitable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misinformation Harms</td>
<td>Making disinformation cheaper and more effective</td>
<td>Quality of service Harms</td>
<td>Alienation</td>
</tr>
<tr>
<td></td>
<td>Facilitating fraud, scams and more targeted manipulation</td>
<td></td>
<td>Increased Labour</td>
</tr>
<tr>
<td></td>
<td>Assisting code generation for cyber attacks, weapons, or malicious use</td>
<td></td>
<td>Service Or Benefit Loss</td>
</tr>
<tr>
<td></td>
<td>Illegitimate surveillance and censorship</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malicious Uses</td>
<td>Anthropomorphising systems to lead to surveillance or unsafe practices</td>
<td>Inter- &amp; intrapersonal Harms</td>
<td>Loss Of Agency, Social Control</td>
</tr>
<tr>
<td></td>
<td>Creating avenues for harm</td>
<td></td>
<td>Technology-Facilitated Violence</td>
</tr>
<tr>
<td></td>
<td>Promoting harmful stereotypes by implying gender or ethnic identity</td>
<td></td>
<td>Diminished Health And Well-being</td>
</tr>
<tr>
<td>Human-Computer Interaction Harms</td>
<td>Environmental harms from operating LMs</td>
<td>Media System/Societal Harms</td>
<td>Information Harms</td>
</tr>
<tr>
<td></td>
<td>Increasing inequality and negative effects on job quality</td>
<td></td>
<td>Cultural Harms</td>
</tr>
<tr>
<td></td>
<td>Undermining creative economies</td>
<td></td>
<td>Political And Civic Harms</td>
</tr>
<tr>
<td></td>
<td>Disparate access to benefits due to hardware, software, skill constraints</td>
<td></td>
<td>Macro Socio-economic Harms</td>
</tr>
<tr>
<td>Automation, access, and environmental harms</td>
<td></td>
<td></td>
<td>Environmental Harms</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Weidinger et al., 2022

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
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

Development

Mitigation

Inference/Generation
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
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
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