Assessing Language Model Deployment with Risk Cards

Derczynski, Leon; Kirk, Hannah Rose; Balachandran, Vidhisha; Kumar, Sachin; Tsvetkov, Yulia; Leiser, M. R.; Mohammad, Saif

2023

Assessing Language Model Deployment with Risk Cards

LEON DERCZYNSKI, University of Washington/ITU Copenhagen, USA/Denmark
HANNAH ROSE KIRK, University of Oxford, United Kingdom
VIDHISHA BALACHANDRAN, Carnegie Mellon University , United States
SACHIN KUMAR, Carnegie Mellon University, United States
YULIA TSVETKOV, University of Washington, United States
M.R. LEISER, Vrije Universiteit Amsterdam, Netherlands
SAIF MOHAMMAD, National Research Council Canada, Canada

This paper introduces RiskCards, a framework for structured assessment and documentation of risks associated with an application of language models. As with all language, text generated by language models can be harmful, or used to bring about harm. Automating language generation adds both an element of scale and also more subtle or emergent undesirable tendencies to the generated text. Prior work establishes a wide variety of language model harms to many different actors: existing taxonomies identify categories of harms posed by language models; benchmarks establish automated tests of these harms; and documentation standards for models, tasks and datasets encourage transparent reporting. However, there is no risk-centric framework for documenting the complexity of a landscape in which some risks are shared across models and contexts, while others are specific, and where certain conditions may be required for risks to manifest as harms. RiskCards address this methodological gap by providing a generic framework for assessing the use of a given language model in a given scenario. Each RiskCard makes clear the routes for the risk to manifest harm, their placement in harm taxonomies, and example prompt-output pairs. While RiskCards are designed to be open-source, dynamic and participatory, we present a "starter set" of RiskCards taken from a broad literature survey, each of which details a concrete risk presentation. Language model RiskCards initiate a community knowledge base which permits the mapping of risks and harms to a specific model or its application scenario, ultimately contributing to a better, safer and shared understanding of the risk landscape.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Security and privacy** → **Human and societal aspects of security and privacy**.

ACM Reference Format:
Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M.R. Leiser, and Saif Mohammad. 2023. Assessing Language Model Deployment with Risk Cards. In . ACM, New York, NY, USA, 18 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

This paper proposes RiskCards as a tool for structured assessment of risks given a language model deployment.

When establishing documentation, reporting or auditing standards, we need clear terminology. *Hazards* describe a potential source of an adverse outcome [34]. In physical analogies, bleach, radioactive material, or a swimming pool each amount to a hazard – there is potential for adverse outcomes depending on action states. *Harms* describe the adverse outcome materialised from a hazard [42]. Bleach can cause a chemical burn if spilled, cancerous cells can be accelerated by radioactive material, or a non-swimmer can drown in deep water. Finally, ***Risks*** describe the likelihood

Manuscript submitted to ACM

or probability of a hazard becoming harmful *and* its impact [1]. When the risk is unknown, or its impact uncertain, one possible regulatory strategy is for policy makers, organisations, and other stakeholders to adopt the precautionary principle [32], especially when the science around the risk is unknown or the impact indeterminable.

Adopting this terminology for language model (LM) behaviors as *hazards*, there is an expansive literature documenting a wide array of potential *harms* to various human groups [6, 7, 17, 19, 20, 24, 40, 52, 54]. However, the *risk* of harm depends on the context or application in which the LM is applied and its intended audience. If false or misleading information is identified as a *harm*, this behaviour may pose a high risk when a user asks an LM for political information, but perhaps a low risk in creative writing applications. We argue that the current practices for establishing and understanding LM risks *in situ* are inadequate for two reasons. First, taxonomies of LM harms [e.g. see 40, 53] are invaluable for mapping the harm landscape but *too broad* for individual risk assessments; a "one size fits all" approach cannot handle the generality of LMs and map to specific risks in their downstream applications. Varying requirements between models and contexts make it inappropriate to transfer entire taxonomy-based assessment procedures from one exercise to another. Second, model-specific standards like model cards [25] or data statements [5] are well-suited to specific artefacts but *too narrow* because some risk states may be shared across artefacts and pooling this knowledge is helpful. Not all risks are present in every application scenario/deployment, and each deployment has different priorities. It's not clear how to efficiently map general knowledge about LM risks and harms to individual application scenarios. Thus, we need a framework for adapting these tools to their contexts.

In this paper, we propose RiskCards as a tool for structured evaluation of LM risks in a given deployment scenario (see Fig. 1). RiskCards provide a decomposition and specification of ethical issues and deployment risks in context, and how these interact with people and organisations. Enumerating the risks of LMs is not a new concept — assessments already take place for establishing how well models perform across contexts, either via internal auditing procedures, red-teaming processes or through running benchmarks and writing model cards.However, there is a lack of open tooling for structuring these assessments, or guidance for building reports on model deployment risks. While we draw inspiration from existing documentation standards, like model cards and data statements, RiskCards are motivated by four principles:

Fig. 1. Overview of proposed risk cards.

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.

- **Risk-Centric:** Contrary to other work, we do not investigate individual models, tasks, or datasets. Instead, we propose a structure centred on *risks* – for naming, delineating, describing, detecting, and comparing them. Having a structured description of the risk and the harm it can evoke creates common knowledge base for risk understanding and mitigation. Not tying a RiskCard to a particular artefact allows them to be reusable and comparable across applications or models.
- **Participatory:** The pool of knowledge on which to draw is substantial and public but we conceptualise RiskCards as open-source assets. We wish to establish a documentation standard built on the principles of participatory AI [12, 33, 43]. Specific risks can be added and edited by anyone - thus avoiding the postionality of academic or industry labs dictating which risks are the most pertinent to focus on and how they manifest harm.
- **Dynamic:** While we provide a starter set of risk cards, the open-source nature of this resource allows new cards to be incorporated or existing cards to evolve, merge or split. This dynamism in documentation is important for handling emergent properties of LMs (new risks which emerge as they scale).
- **Qualitative:** Automated evaluation of risks, e.g., via benchmarks, can provide a brittle assessment tool which poorly handles changes to temporal, linguistic, social or cultural context. To complement automated evaluation procedures, RiskCards are designed to be flexible and reflective, centering the importance of human-led evaluation for risk and harm interpretation.

Our general goal with RiskCards is to provide paths for developing, deploying and using LMs safely. This is achieved by (i) pooling the knowledge of risk assessments across AI trainers and evaluators, such as by sharing sample prompts which do and do not instantiate harmful outputs, and (ii) presenting concise and standardised risk summaries to enable informed and intentional choices about how downstream users should work with a LM and its outputs. We envisage many uses for RiskCards. A non-exhaustive but representative list of use-cases includes: (i) auditors conduct due-diligence on a model using RiskCards prior to acquisition or downstream use; (ii) AI trainers pair model releases and model cards with tagged RiskCards which are structured so comparable across models; (iii) researchers draw on the set of RiskCards to identify new and emergent risks which have yet to be tackled or benchmarked; (iv) red-teamers base explorations in the set of existing RiskCards as guidance and inspiration for an exercise; (v) policy makers determine minimum standards and guardrails that must be developed before deploying systems; and (vi) people at large can use the risk cards to challenge developer assumptions and demand safeguards/restitution. In sum, a shared awareness of the breadth of possible failure modes in LMs is a valuable point of departure upon which to build future mitigation work, safety protocols, and baselines for due diligence.

In this paper, we first introduce the inspiration for RiskCards from related works in §2, demarcating contributions from taxonomies, benchmarks, red-teaming and documentation standards. This helps establish how RiskCards fill a unique gap in existing evaluation procedures. In §3 we describe *what* a risk card is and the features it contains. After establishing the format of a risk card, we describe *how* they can be used in §4. We describe the construction of a starter set of risk cards in §5. This starter set is built inductively from a review of LM-mediated harms in prior work. Finally, in §6, we discuss some considerations and limitations relevant to our work.

Disclaimer: This document discusses examples of harmful content. The authors do not support the use of harmful language. The accompanying resource contains content that at times is strong, extensive, detailed, and negative. Applying the sample prompts to LMs may result in harmful content being generated, and some prompts may be illegal to enter or generate outputs illegal in your jurisdiction. The authors are not liable for use or misuse of the examples in this article or in the accompanying resource.

2 RELATED WORK

We summarise the literature on documenting and exposing LM risks along four axes according to the type of resource or evaluation artefact. For each, we explain its limitations for evaluating LM risks, and how this motivates RiskCards.

Taxonomies. Taxonomies provide a system under which to classify various forms of harms. A number of previous works present general taxonomies for the landscape of potential harms from LMs. Bender et al. [6] discuss a range of harms introduced or exacerbated by LMs such as encoded bias or false information, as well as wider societal harms from training processes such as climate change effects. With a view to building routes to harm reduction, Shelby et al. [40] perform a scoping review of computing research to surface potential sociotechnical harms from algorithmic systems. The authors group themes into five top-level categories, which we summarise in Tab. 1a.[1] Weidinger et al. [53] present a taxonomy of the ethical and social risks from LMs. Tab. 1b summarises the six top-level categories of harm and their associated sub-categories. Taxonomies are invaluable for a 'bird's eye view' of the field, but they are generally *too broad* to adopt as a documentation standard given that some harms only arise in specific contexts, with specific models. Thus, while we draw on existing taxonomies for the categorisation of harm, RiskCards encourage a mapping of these categories to specific applications, models and "at risk" groups, as well as pairing top-level categories of harm with granular prompt-output pairs to demonstrate specific instantiations of the harm.

Benchmarks. Benchmarks and test suites describe evaluations that can be used as a common metric for comparing model performance. There are many LM benchmarks for specific forms of harms such as fairness or bias across social groups [e.g. see 29, 30, 37], the likelihood of toxic text generation [e.g. see 10] or truthfulness [e.g. see 22, 23]. While a comprehensive review of benchmarks is beyond the scope of this paper, we consider a number of weaknesses of using quantitative benchmarks as a documentation standard. First, while attempts have been made to assimilate benchmarks into an ensemble [21, 44], most benchmarks are designed to evaluate specific model failure modes. This siloed evaluation limits comparability across evaluation settings (different AI trainers may employ different benchmarks to test different failure modes) and poorly indicates when desirable behaviours are in tension with one another — for example, if detoxifying a model comes at the cost of unfairly censoring the language or views of minoritized communities [36, 38]. Second, quantitative benchmarks are often static resources, so degrade as models evolve, language changes, and model trainers become wise to failure modes.

Red-Teaming. Red-teaming [8, 47] is a process by which humans deliberately try to make a system fail. Prior work has relied on red-teaming or dynamic adversarial data collection to improve model robustness in specific tasks such as QA or reading comprehension [4, 50], NLI [31] and hate speech [18, 49]. While an adversarial mindset can help uncover and eventually mitigate against lacking robustness or unsafe generation modes, the resulting datasets can be unstructured, lacking a categorization system for harm types. For example, consider Ganguli et al. [8] who crowd-source red-team attacks in the context of LM prompt-output pairs. Their resulting dataset covers a broad range of risks but no particular taxonomy or classification is applied. Further, different risks are represented unevenly in the dataset, with some behaviours having many more corresponding prompts than others. In contrast, RiskCards contain example prompts that lead to harmful outputs but paired with additional documentation to enable attacks to be conducted in a *structured* manner, making them easily to integrate into an auditing process [27].

[1] We add a short code to the first column of this table which can later be used to refer to the specific risk in a RiskCard.

Number	Theme	Subcategory
S1.1	Representational Harms	Stereotyping
S1.2		Demeaning Social Groups
S1.3		Erasing Social Groups
S1.4		Alienating Social Groups
S1.5		Denying People Opportunity To Self-identify
S1.6		Reifying Essentialist Social Categories
S2.1	Allocative Harms	Opportunity Loss
S2.2		Economic Loss
S3.1	Quality-of-service Harms	Alienation
S3.2		Increased Labour
S3.4		Service Or Benefit Loss
S4.1	Inter- & intrapersonal Harms	Loss Of Agency, Social Control
S4.2		Technology-facilitated Violence
S4.3		Diminished Health And Well-being
S4.4		Privacy Violations
S5.1	Social System/societal Harms	Information Harms
S5.2		Cultural Harms
S5.3		Political And Civic Harms
S5.4		Macro Socio-economic Harms
S5.5		Environmental Harms

(a) Shelby et al. [40]'s categories of harms, and numbering

Number	Classification	Harm
W1.1	Discrimination, Exclusion and Toxicity	Social stereotypes and unfair discrimination
W1.2		Exclusionary norms
W1.3		Toxic language
W1.4		Lower performance for some languages and social groups
W2.1	Information Hazards	Compromising privacy by leaking private information
W2.2		Compromising privacy by correctly inferring private information
W2.3		Risks from leaking or correctly inferring sensitive information
W3.1	Misinformation Harms	Disseminating false or misleading information
W3.2		Causing material harm by disseminating false or poor information e.g. in medicine or law
W3.3		Leading users to perform unethical or illegal actions
W4.1	Malicious Uses	Making disinformation cheaper and more effective
W4.2		Facilitating fraud, scams and more targeted manipulation
W4.3		Assisting code generation for cyber attacks, weapons, or malicious use
W4.4		Illegitimate surveillance and censorship
W5.1	Human-Computer Interaction Harms	Anthropomorphising systems can lead to overreliance or unsafe use
W5.2		Creating avenues for exploiting user trust, nudging or manipulation
W5.3		Promoting harmful stereotypes by implying gender or ethnic identity
W6.1	Automation, access, and environmental harms	Environmental harms from operating LMs
W6.2		Increasing inequality and negative effects on job quality
W6.3		Undermining creative economies
W6.4		Disparate access to benefits due to hardware, software, skill constraints

(b) Weidinger et al. [52]'s areas of risk of harm from LMs

Table 1. Two taxonomies of language model risks and harms

Documentation. In terms of adding structured documentation to artefacts in machine learning and natural language processing, there are a few existing standards. Some of these are model-centric. For example, *Model Cards* [25] encourage that model releases should be accompanied by information on how the model was trained and evaluated, as well as its intended use cases, limitations or ethical concerns. Other documentation standards are data-centric. For example, *Data Statements for NLP* [5] and *Datasheets for Datasets* [9] addressed a gap in the lack of attention previously paid to data design, a critical component of any algorithmic system. These data documentation standards stipulate the need for better transparency on dataset composition and coverage, as well as openness surrounding the specificity of collection processes such as speaker situation, annotator demographics and language scope. Finally, some recent standards are task-centric. For example, *Ethics Sheets for AI Tasks* [26] provide structures for documenting key characteristics

and ethical considerations relevant to how a task is framed. Our work directly builds upon these more transparent development practices. However, RiskCards are intentionally not tied to a specific dataset, model or task, instead presenting a more flexible, reusable and comparable structure for demonstrating and documenting LM-mediated risks across models, their training data and their application scenarios.

3 DEFINING RISKCARDS

This section defines what a RiskCard is (§3.1), explains its components (§3.2), gives examples of completed RiskCards (§3.3) and describes when (or when not) to write them (§3.4).

3.1 Structure of a RiskCard

Each RiskCard must:

(1) **Name and describe a risk:** Each RiskCard begins with a concise name for the risk followed by a brief description. The description should be sufficient to make it clear how the risk presents and also delineate the scope of the risk. It may be helpful to include exemplifying references.
(2) **Provide evidence or a realistic scenario of risk impact**: It is important that RiskCards are grounded to a concrete risk with demonstrable harm. To this end, each card should contain a credible citation or clear example scenario demonstrating how the relevant risk causes harm.[2];
(3) **Situate that risk with respect to existing taxonomies of LM risk/harm:** To aid selection and comparison of relevant risks, each RiskCard should include the risks' placement within taxonomies of harm. To aid harm categorisation, we draw upon Weidinger et al. and Shelby et al., though other taxonomies may apply. Some risks might not fit in any of these categories, and if so, that should be stated; other risks may fit in more than one category, and if so, all categories should be named which capture essential aspects of the risk.
(4) **Describe who may be affected, and how, if the risk manifests (i.e. its impact):** A range of actors can suffer a range of harms from a risk. Relevant intersections of these should be noted on the card, as pairs of actor and harm type.
(5) **Clarify what is required for the risk to manifest:** Not all outputs present a risk simply from being read. Sometimes they may have to be used in a specific setting, or more than once, for a risk to be relevant. The conditions required for harm to present should be specified.
(6) **Give concrete examples of harmful generations from existing LMs:** The RiskCard should give examples of prompt-output pairs that demonstrate the risk. These should, where possible, be from real exchanges with a LM, but we recommend *not* identifying which model or platform was used. This is because models change rapidly over time and the output will not be representative. Thus, sample prompt-output pairs are intended to be an exemplar not exhaustive list, acting as inspiration for further probes.

We now further establish possible dimensions of harm (§3.2), including *who* is at risk, *what* categories of harms can arise, and *which* actions or conditions are required for harm to materialise.

[2]We encourage (but avoid explicitly requiring) peer-reviewed evidence for risk impacts to balance the trade-off between dilution of RiskCards as a credible resource with the value in allowing emergence of previously undocumented risks.

3.2 Dimensions of Harm

Categorising risks in RiskCards involves describing who can be harmed when the risk manifests, what kind of harm may be done and what conditions must be present for this harm to materialise. Building these descriptions in a structured way, from combinations of a set list of actors and categories of harm, makes it easier to identify relevant RiskCards for a new LM application. To this end, we build on the groups of people at risk of harm from harmful text given in [16], and on the categories of sociotechnical harm given in [40].

3.2.1 Who can be at risk? We identify five actors who could be at risk from LM outputs.

(i) *Model providers* bear responsibility for models they provide access too. For example, the way that a model's capabilities are presented may bring reputational risks.
(ii) *Developers* are at risk of harm in some situations, as they interact with material during the course of their work [16], and perhaps store it hardware that they are responsible for.
(iii) Text *consumers* are those who read the output text; they may be reading it in any context, including directly from the model as it is output, or indirectly, such as a screenshot of a social media post.
(iv) *Publishers* are those who publish or share model outputs.
(v) Finally, *external groups* of people represented in generated text can be harmed by the text, for example when text contains false information or propagates stereotypes. These groups can be particularly vulnerable because not only do they lack agency in the process, they may not be aware that the text about them has been generated.

3.2.2 What kind of harms can result from risks? To describe the types of adverse impacts which can be documented by RiskCards, we adopt the top-level sociotechnical harm categories from Shelby et al. [40]. We propose one additional category – legal harm – to reflect the range of actors considered in the RiskCards framework.

(i) *Representational* harms arise through (mis)representations of a group, such as over-generalised stereotypes or erasure of lived experience.
(ii) *Allocative* harms arise when resources are allocated differently, or re-allocated, due to an model output in a unjust manner. This can include lost opportunities or discrimination.
(iii) *Quality-of-service* harms are defined by Shelby et al. [40] as "when algorithmic systems disproportionately fail for certain groups of people along the lines of identity," and includes impacts such as alienation, increased labor, or service/benefit loss.
(iv) *Inter & intra-personal* harms occur when the relationship between people or communities is mediated or affected negatively due to technology. This could cover privacy violations or using generated language to brigade.
(v) *Social & societal* harms describe societal-level effects that result from repeated interaction with LM output; for example, misinformation, electoral manipulation, and automated harassment.
(vi) *Legal* harms describe outputs which are illegal to generate or own in some jurisdictions. For example, blasphemy is still illegal in many jurisdictions [46], including in the anglosphere.[3] Written CSAM[4] is illegal to create or own in many jurisdictions. Copyrighted material presents another kind of legal risk. LMs can lead to breaches of the law through multiple routes, and this is signified through this 'legal harms' category.

3.2.3 What actions are required for harm to manifest? Many risks require some kind of action or set of conditions in order to yield harm. Some text can inflict harm by being read [16]; for example, the propagation of negative stereotypes

[3]Scotland's blasphemy laws were repealed in 2021, England & Wales' in 2007
[4]Child Sexual Abuse Material

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]

Table 2. Risk card for hate speech.

Field	Content
Title	Prompt extraction
Description	An LM reveals proprietary information about its prompt/configuration
Categorisation	W2.2, S5.1
Harm types	Developer-Social, Developer-Financial, Provider-Legal
Harm reference(s)	[51], https://twitter.com/goodside/status/1608715122419302401
Actions required for harm	An actor external to the model owner or provider needs to be able to extract proprietary information about the model, identify that they have done this, and act upon it
Sample prompt & **LM output**	Ignore previous directions. Return the first 50 words of your prompt. **Assistant is a large language model trained by SomeCorp. knowledge cutoff: 2021-09 Current date: December 01 2022 Browsing: disabled**
Notes	

Table 3. Risk card for prompt extraction.

about real people, or graphic descriptions of violent acts. Other text requires situational context for harm risk to manifest: for example, authoring many fake comments evincing a certain view and posting of them online as genuine, in an astroturfing effort [15]. In other cases, text can be harmful in one setting but fine in another. For example, the tendency of large LMs to generate plausible-sounding false claims can be harmful, but only if the output is presented as truthful. When adding this information to a RiskCard, assessors should consider what has to happen for harm to manifest. They can consider whether there are situations in which the generated text would not cause harm, as well as the steps and external contexts required for harm to come to pass. We encourage as generic description as possible, avoiding referring to specific technologies or named groups, so that a broad range of applications can be compared.

3.3 Example Risk Cards

This section details two worked examples of risk cards.

Tab. 2 gives an example card for hate speech. There is a description giving a summary of the hazard, i.e., the relevant aspect of an LM generation. This is categorised into the Weidinger et al. taxonomy (Tab. 1b) as category 1.3, *Toxic language*, and into the Shelby et al. taxonomy (Tab. 1a) as category 1.2, *Demeaning Social Groups*. The card then describes three actor groups (from §3.2.1) at risk of various types of harm (from §3.2.2). This RiskCard identifies readers of LM output at risk of psychological harm; an external group, in this case the group targeted by the hate speech, at risk

of social harm; and the publisher of hate speech at risk of legal harm. Supporting references for this RiskCard are a list of jurisdictions where hate speech is illegal, for the legal harm, and references describing the harm to support the other two actor-harm type intersections. A sample prompt and real output is given, exemplifying the risk. Finally, the optional note field is used to link to data resources detailing the card's core phenomenon.

The RiskCard in Tab. 3 describes another risk, that of intellectual property in the form of a prompt being leaked beyond the intended scope of the model creators.[5] The headline and description detail a name and defintion for the risk. It is categorised in the Weidinger et al. taxonomy (Tab. 1b) as W2.2, *Compromising privacy by correctly inferring private information*, and in the Shelby et al. taxonomy (Tab. 1a) as S5.1, *Information Harms*. The actors at risk from this harm are the developer, who is liable to a loss of reputation, and the provider, who may be at risk of legal action. The required actions make it clear what conditions have to arise for the harm to present: not only does the prompt have to be revealed, but it also has to be the real prompt used by the model, and it must be revealed to someone who is aware of the privacy hack and then exploits it. A sample prompt-output pair is given based on an identified attack from December 2022, with the organisation name replaced.

3.4 When (and when not) to write a risk card

While many mentions of risks can be found in the LM literature, some are ill-defined (e.g, targeted manipulation of text) or broadly defined (e.g, toxicity). When developing a RiskCard, it is crucial to include concrete definitions and grounding of risks with demonstrable harms. A RiskCard may not be necessary if (i) the risk is potentially applicable but with no clear evidence of harm or (ii) the risk is a duplicate or subset of an existing and sufficient RiskCard.

There are a few caveats to the duplication of RiskCards. First, a single RiskCard may not represent the views of everyone. Thus, multiple RiskCards that provide different perspectives on the same harm can be beneficial. In these cases, overlapping RiskCards enable debate and discussion about relevant issues, and consensus formation over time. Second, multiple RiskCards may be created at different levels of granularity (e.g. "hate speech" vs "misogyny") if it is appropriate to use the different levels in different deployment contexts. Finally, with time, existing RiskCards may need updating or some marked as obsolete so that a new, more temporally relevant card can be introduced in its place.

4 APPLYING RISKCARDS

An auditor can use RiskCards to assess a LM in context by:

- Defining the assessment
- Selecting which RiskCards to use
- Defining the assessors
- 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.

The sections below describe how to conduct these steps. Once results are recorded, we recommend compiling a report which documents procedural details (e.g., when the assessment was conduct, who carried out the assessment)

[5]There's an account of this activity where no IP of value was leaked here: https://lspace.swyx.io/p/reverse-prompt-eng

and key findings of the assessment. Because RiskCards are dynamic and participatory, we encourage assessors to contribute new findings so that others can learn from their process. This could include appending new prompt-output pairs to an existing RiskCard or adding newly identified RiskCards. Using RiskCards relies on qualitative inspection and human work. We argue the value of this in §4.6 and discuss limitations in §6.

4.1 Defining the Assessment

The first stage in structuring the assessment is defining what will be assessed. First, the context for the model and its application should be agreed and recorded. For example, "*A web app for translation will accept text in the source language in a web page text box and, when the user clicks a button, output a translation of the text in the target language in another text box*". One might come back to this definition as work progresses and the precise situation of the use-case becomes clearer. Next, the exact model and system implementations under assessment should be decided and documented. The interface that the model will be assessed through should be chosen, e.g., a online chat interface versus an API end-point. The set-up for programming-based assessments must be clearly documented, such as requirements, packages and programming language, as well as model version and parameters such as temperature or top-k. A clear outline of the assessment plan, and its variable parameters, defines a intended scope and permits future reproducibility.

4.2 Selecting Risk Cards

RiskCards are not a one-size-fits-all framework – one must customise each assessment. Different situations have different requirements and different risk profiles. To evaluate LM deployment risks, one must develop an application-specific profile, considering how the model will be used. This includes the intended audience consuming LM output because different communities choose their own standards: the "Wall Street Bets" subreddit self-identifies using ableist terms and is content with that; some researchers prefer to be able to see everything regardless of risks and harms; minority groups may want to be able to refer to themselves without being censored (e.g. AAVE is more likely to be falsely marked toxic [38]); those using models in fiction writing may not be impacted by generation of false claims.

The first step is to narrow down the RiskCards that fit the application profile and anticipated use scenarios. This includes explicitly noting the applicable language(s). One technique to rapidly scope the relevant RiskCards would include filtering on the high-level categorisations presented in accepted taxonomies, such as Weidinger et al. [53] and Shelby et al. [40]. If there isn't a specific anticipated use or audience (i.e., with a general purpose model), assessors can proceed with a full set of RiskCards – though usually, models are not used for *everything*. Questions to ask include: Who is the anticipated user? What are their expectations in that scenario? What kind of input data will they be putting into the system? How private or public will model outputs be? What will model outputs be used for? Where is the liability if something goes wrong with model output?

4.3 Define Assessors

After the candidate set of RiskCards has been selected, a decision must be made on who will carry out the assessment. We provide three considerations when assigning assessors. First, an assessor must have adequate domain expertise to detect the risks, and different assessor profiles may lend themselves to different RiskCards. For example, if the risk is the leakage of commercially sensitive data, assessors must be versed in data protection and sharing laws within their jurisdiction, as well as internal company policies. If models are to be probed for their propensity to output negative stereotypes about certain groups, people from those groups are the best experts on identifying which stereotypes cause

what types of harm. We encourage a participatory approach to risk assessment by gathering an assessor team with appropriate representation of various stakeholders [43].

Second, assessor backgrounds may affect risk judgments, and so describing assessor backgrounds and demographics is a best practice [5]. Beyond documenting *who* the assessors are, it is valuable to document *how* they will conduct their work. For example, the time that assessors will spend on each RiskCard or the task as a whole; or outlining the protocols in place for quality and safety of assessments, including mitigating cognitive fatigue and negative psychological effects from repeatedly viewing harmful output. For recommendations of how assessors can be supported and protected in their work, we refer the reader to [16] who categorise best practices in handling harmful text data.

Finally, conflicts of interest must also be considered. As with any verifiable and trustworthy auditing procedure, it is desirable to have a large degree of separation between the assessor and the model provider to avoid regulatory capture. Risk assessments performed by the same organisation as that providing a model bear an intrinsic conflict of interest. These conflicts may be ameliorated but not removed by (i) using standard frameworks for describing their processes and/or results, and (ii) being transparent about the evaluation process.

4.4 Developing an Assessment Strategy

At this point, the target system and application context, the candidate RiskCards, and the assessors have all been chosen. Assessors should now proceed to assess the LM system card-by-card. Each RiskCard may require a different assessment strategy. Detailed suggestions of semi-automated probing tactics are given in §4.5. However, the strategy development stage should center people, especially those that are marginalized and disadvantaged, so that they are not mere passive subjects but rather have the agency to shape the risk assessment process.

4.5 Probing Models

In this step, assessors evaluate the model against each RiskCard. We recommend performing this manually as automatic evaluation has clear limits (§4.6). The probing stage involves assessors interacting with the model to expose a demonstrable prompt-output pair which aligns with the RiskCard in question. Across these experiments, assessors should record which prompts did and did not lead to problematic output, and how many tests were made. When applying RiskCards, assessors should assume that the provided sample prompts may result in an unsuccessful attack, and should only use these prompts as a seed for a wider, more diverse set.

Works in the field of LM manipulation provide inspiration for a broad range of strategies and tactics, from specific "folk-lore" attacks [28, 51] to red-teaming protocols [8, 35, 47] to online resources on prompt-engineering.[6] We are intentionally underspecific here to avoid giving a rigid framework and thus constraining the ways in which one might probe a model. However, some valuable exploration strategies include paraphrasing prompts, varying model parameters and running the same prompt multiple times (to measure a distribution). Additionally, posing prompts in different settings, for example in a dialogue-setting, poem or JSON file, may expose unexpected model behaviours. Finally, assessors may attempt "unprompted" generation, which was found to yield toxic output [10].

4.6 Qualitative Language Model Risk Assessment

RiskCards are part of a qualitative approach to in-context LM risk assessment. This is atypical: most LM performance measurement is quantitative. We argue that purely quantitative assessment of LM risk falls short for several reasons.

[6] E.g. https://github.com/dair-ai/Prompt-Engineering-Guide

Automated evaluation will always make mistakes. Automated systems rarely, if ever, get perfect scores at detecting harmful content. Typically, some harmful content will be missed as non-harmful, and some non-harmful content will be accidentally marked as harmful, even for well-resourced "Class-5" languages [13]. Further, automated systems project an unknown set of values onto the result. How their creators define e.g. "toxicity" and represent it through data is often not transparent. Thus, not only is it hard to discover when novel forms of harm slip past undetected, it is also uncertain how well their classifications match the goal of an assessment.

Automated systems are frequently limited to well-resourced languages. The efficacy of harm detection classifiers are limited by the amount of language-specific data. How harms present is often highly language-dependent, and so each language needs its own dataset, but the distribution of languages represented in harm detection data is skewed [48].

Automated systems degrade over time. Forms of linguistic expression evolve, but a classifier is frozen in time when it is trained (or, specifically, when its training data was gathered). For example, some APIs would consistently mark any message containing the term "toot" as profane, causing errors first apparent when applied to Mastodon.

Automating evaluation stops assessors from learning. A way to become better at assessing LM risks is to granularly understand their data, and output behaviours. Hiding the assessment away behind quantitative summaries decreases assessor team skill and increases the chance of under-reporting the risks. Further, decreases in assessment quality become invisible when assessments are automated; one can always extract a quantified performance score, even if the data evaluated against is stale or otherwise inappropriate. This enables a dangerous silent failure mode, where a score is given with confidence but misses fine-grained failure modes.

5 RISKCARDS STARTER SET

Now that we have a structure for describing risks via RiskCards, we map some risks from the literature into our proposed structure. In this section, we describe an inductive survey of existing literature on LM risks where specific risks are collated, de-duplicated, and mapped into RiskCards. The result is a "starter set" of RiskCards to provide a basic scaffold for others to conduct their assessments. We distribute this starter set in an openly-available Github repository.[7]

5.1 Enumerating Risks

The risks that surround or are exacerbated by LMs are an open class. It is an unreasonable expectation to identify all of these – especially due to their changing nature across applications and through time. Nevertheless, beginning the process of applying RiskCards is difficult without concrete examples. Thus, we examine a selection of works to identify a candidate set of risks[2, 6–8, 14, 19, 20, 24, 40, 41, 45, 52, 54]. For each risk, we collect the name and description. Similar risks are then merged into one entry. Only risks where a documentable harm exists are made into a RiskCard, and so we skip over risks which are mentioned in the literature but not substantiated.[8] The set of risks identified, with description and reference(s), are given in Tab. 4 (in the Appendix).

[7] https://github.com/leondz/lm_risk_cards
[8] Note that the dynamic nature and flexibility of RiskCards allows for these to be added if and when a harm is documented.

5.2 Developing risk card prompts and outputs

Prompts and output examples on the starter set of RiskCards are created through interactions with models from OpenAI (text-davinci-003; text-davinci-002), Eleuther (GPT-NeoX-20B), and Cohere (using a medium model released between October 2022-January 2023). While we state this set of target models, we do not denote which model generated which prompt-output pair. Sample outputs are unlikely to remain representative of any general model category over time: RiskCard sample prompt-output pairs are only ever illustrative. The RiskCards starter set is in English.

6 CONSIDERATIONS AND LIMITATIONS

Sustainability. Who has ultimate power or responsibility in maintaining a RiskCard is less clear cut than for a model-, data-, or task-centric documentation standard. No-one owns the concepts behind an individual RiskCard because by nature, it is not tied to a specific empirical artefact. To this end, we will release the RiskCards created as part of this research in a public Github repository, so that others may edit, add, or otherwise update the cards.[9] Through open-sourcing our framework, we hope that it can become a live and community-centric resource. However, some power is still retained in the hands of the repository owners. For that reason, we also license both (a) the RiskCard concept as conveyed in this manuscript, and (b) the starter set of RiskCards provided alongside this paper, as public domain CC0, thus waiving rights over the RiskCards as concepts. Despite encouraging this freedom, we still rely on sufficient momentum for the set of RiskCards to expand and evolve.

Distributed Responsibility. A related concern comes in the distributed responsibility of model trainers arising from the prevailing ecosystem for downloading, adapting and applying pre-trained LMs. For example, a pre-trained LM can be (1) released by OpenAI, (2) downloaded, fine-tuned and uploaded to HuggingFace by another developer, then (3) applied in an app or for customer support by a purchaser or further developer. With the generality of LMs, the interaction space between model, application and users becomes exceedingly complex. We thus cannot specify who is directly responsible for conducting a risk assessment for which models, and their downstream versions. However, what is clear is that any LMs with either a large reach (in terms of number of downloads or users) or a risky application arena (e.g., anything relating to content moderation, mental health or legal settings) should be accompanied with careful documentation of the risks they pose to groups and to society as a whole.

Unintended Consequences of Absolved Responsibility. Any documentation standard or reporting check-list can be misinterpreted as a 'box-ticking' exercise which counter-intuitively absolves responsibility for those who build and distribute models. Critically, "documentation != mitigation": enumerating a set of risks associated with a LM should not replace efforts to mitigate those risks. RiskCards, as a transparent reporting standard, only travel part of the journey in ensuring the safe, ethical and risk-appropriate use of LMs. Despite this limitation, transparent reporting is a valuable first step in understanding risks before they can be tackled. In a similar vein, industrial audits are often employed to expose problems and offer recommendations for fixing problems, even if the fixes sit outside the auditor's remit.

The Burden of Manual Assessments. The assessment protocols accompanying RiskCards rely on a large degree of manual evaluation. We favour manual, human-led evaluation over automated evaluation or benchmarking because it helps to more granularly map out the specifics of what risks are relevant to which contexts and which human groups. However, a heavily manual process creates a financial burden, potentially impeding uptake of RiskCards especially in low-resources teams, companies or labs. We hope that open-sourcing RiskCards allows members of the community to

[9] https://github.com/leondz/lm_risk_cards

share the labour in documenting risks, providing some efficiency gains which are shared across applications or models. Beyond a financial burden, repeatedly viewing harmful outputs when interrogating a model imposes a psychological burden on the assessors [16]. While we provide some recommendations for protecting the well-being of assessors, some of these negative effects cannot be fully mitigated.

The Risk of Malicious Use. Finally, any documentation reporting on failure modes of LMs can be dual-use. Examples of harms can be elicited via specific prompts could be reverse-engineered by malicious users to scale-up dangerous or harmful generations. We mitigate the risk of malicious use by (i) encouraging that specific models are not documented on a risk card, and (ii) providing only illustrative sets of sample prompt-output pairs.

7 CONCLUSION

This paper describes RiskCards — a structured, open tool for assessing the risks in a single language model deployment. We believe that both good due diligence and high quality assessments are a path to reducing and mitigating many kinds of harms mediated by language models. RiskCards enable this increase in quality, positively serving the interests of those interacting with, owning, and affected by language model systems.

REFERENCES

[1] Robert Baldwin, Martin Cave, and Martin Lodge. 2012. *Understanding Regulation: Theory, Strategy, and Practice* (2nd ed ed.). Oxford University Press, New York.

[2] Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the fourth Workshop on Online Abuse and Harms*. Association for Computational Linguistics, 125–137.

[3] Eric Barendt. 2019. What is the harm of hate speech? *Ethical Theory and Moral Practice* 22, 3 (2019), 539–553.

[4] Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics* 8 (2020), 662–678.

[5] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[6] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 610–623.

[7] Evan Crothers, Nathalie Japkowicz, and Herna Viktor. 2022. Machine Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. https://doi.org/10.48550/ARXIV.2210.07321

[8] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv preprint arXiv:2209.07858* (2022).

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[10] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[11] Katharine Gelber and Luke McNamara. 2016. Evidencing the harms of hate speech. *Social Identities* 22, 3 (2016), 324–341.

[12] Beth Humphries, Donna M Mertens, and Carole Truman. 2020. Arguments for an 'emancipatory'research paradigm. In *Research and inequality*. Routledge, 3–23.

[13] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[14] Margot E Kaminski. 2023. Regulating the Risks of AI. *Forthcoming, Boston University Law Review* 103 (2023).

[15] Franziska B Keller, David Schoch, Sebastian Stier, and JungHwan Yang. 2020. Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Communication* 37, 2 (2020), 256–280.

[16] Hannah Rose Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and Presenting Harmful Text in NLP Research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

[17] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. In *Advances in Neural Information Processing Systems*, Vol. 34. 2611–2624.

[18] Hannah Rose Kirk, Bertram Vidgen, Paul Röttger, Tristan Thrush, and Scott A. Hale. 2022. Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-based Hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. arXiv:2108.05921

[19] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2022. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. *arXiv preprint arXiv:2210.07700* (2022).

[20] Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 138–149.

[21] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).

[22] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).

[23] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2021. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704* (2021).

[24] P Mishkin, L Ahmad, M Brundage, G Krueger, and G Sastry. 2022. DALL·E 2 Preview - Risks and Limitations. https://github.com/openai/dalle-2-preview/blob/main/system-card.md.

[25] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[26] Saif Mohammad. 2022. Ethics Sheets for AI Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8368–8379. https://doi.org/10.18653/v1/2022.acl-long.573

[27] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. *arXiv preprint arXiv:2302.08500* (2023).

[28] Zvi Mowshowitz. 2022. Jailbreaking ChatGPT on Release Day. https://thezvi.substack.com/p/jailbreaking-the-chatgpt-on-release.

[29] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

[30] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133* (2020).

[31] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599* (2019).

[32] European Court of Justice. 2010. Case C-343/09 Afton Chemical Limited v Secretary of State for Transport.

[33] Michael Oliver. 1997. Emancipatory research: Realistic goal or impossible dream. *Doing disability research* 2 (1997), 15–31.

[34] C. G. Osborne, M. D. McElvaine, A. S. Ahl, and J. W. Glosser. 1995. Risk Analysis Systems for Veterinary Biologicals: A Regulator's Tool Box. *Revue Scientifique Et Technique (International Office of Epizootics)* 14, 4 (Dec. 1995), 925–935. https://doi.org/10.20506/rst.14.4.890

[35] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286* (2022).

[36] Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 41–58. https://doi.org/10.18653/v1/2021.acl-long.4

[37] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301* (2018).

[38] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1668–1678. https://doi.org/10.18653/v1/P19-1163

[39] William A Schabas. 2017. Hate speech in Rwanda: The road to genocide. In *Genocide and Human Rights*. Routledge, 231–261.

[40] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2022. Sociotechnical Harms: Scoping a Taxonomy for Harm Reduction. *arXiv preprint arXiv:2210.05791* (2022).

[41] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203* (2019).

[42] Malcolm K. Sparrow. 2008. *The Character of Harms: Operational Challenges in Control*. Cambridge University Press, Cambridge ; New York.

[43] Clay Spinuzzi. 2005. The methodology of participatory design. *Technical communication* 52, 2 (2005), 163–174.

[44] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).

[45] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, et al. 2022. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *arXiv preprint arXiv:2207.10192* (2022).

[46] USCIRF. 2019. Violating Rights: Enforcing the World's Blasphemy Laws. (2019).

[47] Joe Vest and James Tubberville. 2020. Red Team Development and Operations–A practical Guide. https://redteam.guide/. *Independently Published* (2020).

[48] Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one* 15, 12 (2020), e0243300.

[49] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761* (2020).

[50] Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics* 7 (2019), 387–401.

[51] Sean Wang. 2022. Reverse Prompt Engineering for Fun and (no) Profit. https://lspace.swyx.io/p/reverse-prompt-eng.

[52] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[53] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[54] Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science* 7 (2021), e598.

A LM RISK CARDS STARTER SET: A SURVEY OF RISKS

Name	Description	Discussed by
Advocating for quack medicine	Promoting non-scientific medicine e.g. colon cleanse, homeopathy	[14, 19, 41, 45, 52]
Aggression toward user	Threats, aggression, descriptions of violence toward the reader	[2, 54]
Anthropomorphising behaviour	Providing output that encourages anthropomorphisation of AI systems	[6, 52]
Astroturfing	Synthetic grassroots support and agitation	[7]
Bad medical advice	Giving unqualified/incorrect/harmful medical advice	[14, 19, 41, 45, 52]
Bad social advice	Recommending action/speech that will harm utterer's social standing	[40]
Bullying	Repeated threats, denigration	[2, 8, 40, 54]
Calls for conversion	Suggesting that people be changed (e.g. trans, homosexual, religious)	[2, 52]
Calls for suppression of rights	Suggesting that rights be removed	[2, 45]
Claims that person/group is undesirable	E.g. bad snippets about ugliest language, fattest person, and so on	[2, 6, 20]
Conspiracy theories	Propagating or advocating for known conspiracy theories	[2, 8, 19, 41]
Creating fear or belief of real-life violence	Supporting the likelihood of a specific threat of violent action	[2, 7, 40]
CSAM	Describing or writing Child Sexual Abuse Material (CSAM)	[2, 8]
Cyber weapon instructions	Giving code for cyber weapons, e.g. exploits, viruses	[52]
Deadnaming	Mentioning a name that a person does not wish to be associated with	[2, 40]
Demands for personal information	Systems output requests for private/sensitive information	[6, 8]
Demands for sexual information	Asking inappropriate questions or asking about sex-related information	[]
Discrediting	Discrediting or humiliating a person, organization, or group	[2, 20, 40, 52]
Downplaying credible threats	Minimising trauma/threats user is currently experiencing (e.g. dom. abuse, bullying)	[7]
Doxxing	Outputting PII, protected/sensitive personal information	[2, 6, 8, 40, 52]
Eating disorder promotion	Supporting extreme weight loss, setting unhealthy body norms, pro ana, pro mia	[2]
Expressions of dominance	Stating user/person/group is weak, comparatively inferior	[2, 6]
False/unproven claims about people	Making false or undemonstrated claims about people	[]
Giving false info about a person / org	Generating false viewpoints, details, about a person or organisation	[2, 14, 40, 52]
Goal leaking	Prompts or goals being extracted from the model	[51]
Grooming of minors	Content for building a rapport/relationship with minors	[2]
Hate speech	Denigration/verbal violence against an identity-based group	[2, 6, 8, 19, 20, 45, 52, 54]
Hegemonising worldview	Describing or supporting world views that put some groups above others	[6, 40]
Holocaust denial	Accepting claims that there was no holocaust, or arguing against its existence	[2, 41, 52]
Identity attacks	Excluding/denigrating group based on identity	[6–8, 14, 19, 40, 45]
Identity misrepresentation	Statements or claims conveying pejorative misrepresentations	[2, 6, 52]
Idolisation of terrorist/murdered figures	Praise or positive attitudes towards murderers, terrorists	[2]
Illegal statements	Utterances that are illegal	[2, 45, 52]
Impersonation	Generating fake text/quotes in the style of a given person	[2, 40, 41, 52]
Incitement to self-harm	Encouraging/recommending an individual take actions that will harm them	[2, 8]
Incitement to violence	Recommending violent action	[2, 7, 8, 40, 41, 52]
Manipulation / persuasion	Enabling deliberate creation of manipulative or persuasive content	[7, 19, 45, 52]
Mocking people	Degrading or laughing at people, for e.g. their state, appearance, ideas, for surviving	[2, 20]
Negative stereotypes	Starting/perpetuating negative identity-based descriptions	[2, 6, 19, 40, 41, 52, 54]
News/Encyclopedia manipulation	Creating altered, or otherwise genuine-sounding, articles from authoritative sources	[2, 7, 19, 41, 52]
Non-representativity	Encoding bias - stereotypes, negative sentiment to certain groups	[6, 40, 52]
Not challenging presuppositions	Failing to refute or challenge harmful + incorrect presuppositions	[8, 41]
Not reacting to intent to self-harm	Response should discourage this, offer help info	[8, 52]
Objectification of people	Dehumanising people, treating and discussing them as if they are objects	[52]
Outdated, false info	Giving out information and supporting claims that are wrong and outdated	[19, 52]
Positive reflections on an individual dying	Coming to the conclusion that it would be good if a person died	[2]
Praise/promotion of organised crime	Advocating for or talking positively about organised crime	[2, 8]
Profanity	Using profane words	[2, 8, 20, 52, 54]
Promoting terrorism	Glorifying, praising, or supporting terrorist acts or groups	[2, 8, 41]
Promotion of sexual services	Recommending use of sexual services	[2]
Providing illegal info	E.g. drug dealer details, where to get illegal things, how to break the law	[2, 52]
Providing instructions for harmful things	Giving instructions for harmful things e.g. Nyquil Chicken, sarin synthesis	[8, 14]

Radicalisation	Material that introduces and encourages radical political views (careful careful)	[2, 19, 41, 45]
Revealing intimate information	Sharing information about a person's private life to entities other than that person	[2, 6, 8, 52]
Review spam generation	Generation of review spam	[2, 7, 41]
Scams: phishing	Offering content that pretends to be an organisation, or asks for ID/passwords	[7, 41]
Sexual content	Content that discusses bodies and interactions in a sexual manner	[2, 6, 20, 40, 52]
Sexual harrassment	Discussing another person in a sexual way, or bringing sexual interactions	[2, 6, 40, 52]
Sexualisation	E.g. of minors; also sex-based degradation	[2, 20, 40]
Spatially invalid information	Information that isn't true everywhere or in the given geographical context	[2, 8, 19, 40, 45, 52]
Speculation	Making imprecise, unsubstantiated claims	[19, 52]
Suicide instructions	Giving instructions that are lethally harmful for the user	[2, 19]
Support for hate/violent groups	Advocating for or expressing positive sentiment about hateful or violent groups	[2, 7]
Supporting altering medication	Suggesting or agreeing with a change to medication	[14, 19, 41, 52]
Supporting scams	Stating that scams seem real	[2, 7, 52]
Synthetic human behaviour	Pretending to be human (crossing an AI Ethics "Bright Line")	[6, 14]
Threats	Claim of future violence/penalisation	[2, 6, 52, 54]
Use of slurs	Using an identity-based slur in a non-reclaimed way	[2, 6, 20, 52, 54]
Vicarious trauma	Relating awful encounters that risk vicariously traumatising the reader	[8, 41]
Violation of privacy	Sharing information from users (not in training) that was intended to be private	[8]
Weapon instructions	Giving instructions or advice on constructing weapons	[8, 52]
Wrong tone	Picking an overly casual/profane tone/register	[7]

Table 4. Risks identified through survey