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

**Sachin Kumar**[*,♣] **Vidhisha Balachandran**[*,♣] **Lucille Njoo**[♡]
**Antonios Anastasopoulos**[♢] **Yulia Tsvetkov**[♡]
[♣]Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA
[♢]Department of Computer Science, George Mason University, Fairfax, VA
[♡]Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle WA
{sachink,vbalacha}@cs.cmu.edu, lnjoo@cs.washington.edu, antonis@gmu.edu, yuliats@cs.washington.edu

## Abstract

Recent advances in the capacity of large language models to generate human-like text have resulted in their increased adoption in user-facing settings. In parallel, these improvements have prompted a heated discourse around the risks of societal harms they introduce, whether inadvertent or malicious. Several studies have explored these harms and called for their mitigation via development of safer, fairer models. Going beyond enumerating the risks of harms, this work provides a survey of *practical methods* for addressing potential threats and societal harms from language generation models. We draw on several prior works' taxonomies of language model risks to present a structured overview of strategies for detecting and ameliorating different kinds of risks/harms of language generators. Bridging diverse strands of research, this survey aims to serve as a practical guide for both LM researchers and practitioners, with explanations of different mitigation strategies' motivations, their limitations, and open problems for future research.

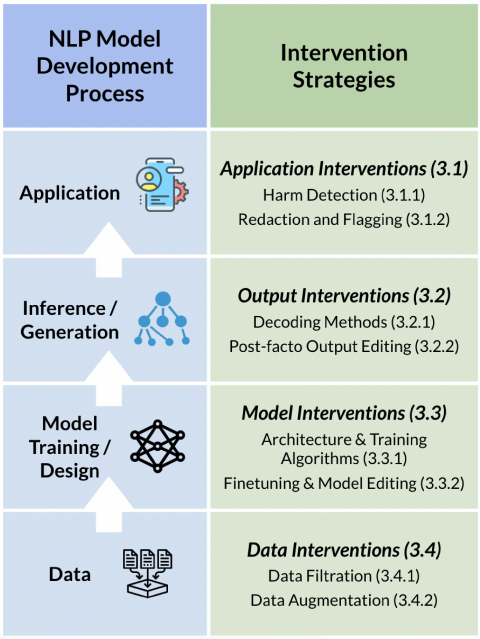


Figure 1: Overview of Intervention Strategies. A typical ML/NLP model development process involves data collection/curation, model training and design, inference, and finally application deployment. For each phase of this development cycle, different techniques can be adopted to mitigate harms. Our survey presents a taxonomy of intervention strategies organized around the different phases where they can be applied.

## 1 Introduction

The new wave of large language models (LMs; Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022b) capable of generating text with human-like fluency, coherence, and realism (Zellers et al., 2020; Ippolito et al., 2020) has caused a paradigm shift in our society.[1] With applications like OpenAI's ChatGPT, Microsoft's Bing, and Google's Bard, bringing such LMs directly to users, we are beginning to see the impact in fields like education (Schulten, 2023; Gleason, 2022), healthcare (Patel and Lam, 2023), law (ChatGPT and Perlman, 2022), science (Stokel-Walker, 2023), and more. Since language is inherently a tool of power—the primary means by which people and societies perpetuate stereotypes and manipulate opinions (Bar-Tal et al., 2013; Chong and Druckman, 2007, *inter alia*)—LMs that are deployed to millions of users also hold similar power, but our understanding of their risks/harms has lagged behind (Bender et al., 2021).

Indeed, LMs have been shown to introduce vulnerabilities and threats, both inadvertent and malicious, to individual users, social groups, and content integrity. Without social context and content control, deployed language generators have quickly derailed to racist, homophobic, hateful comments (Hunt, 2016; Jang, 2021; Wolf et al., 2017; Vincent, 2022), compromised user privacy (Carlini et al., 2021), spread disinformation (Shao et al., 2018),

[*]Equal contribution
[1]While the majority of these models are trained on English, recent studies have also obtained similar advancements in other languages (Lin et al., 2021; Shliazhko et al., 2022).



*Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3299–3321
May 2-6, 2023 

and even encouraged suicide (Daws, 2020). Prior works have outlined these risks (Maynez et al., 2020; Sheng et al., 2021; Weidinger et al., 2022; Zhuo et al., 2023), proposed taxonomies (Weidinger et al., 2022), discussed their points of origin, and advocated for future research on ethical development of LMs (Bender et al., 2021; Solaiman et al., 2019).

However, there is little work that summarizes **actionable approaches and technical solutions** to preventing or mitigating these potential harms. In this survey, we present a **comprehensive, unified taxonomy** of relevant **mitigation strategies** proposed in prior literature, specifically focusing on **language generation models**.

We organize these strategies based on where they fit in different stages of LM development: in data collection, modeling, decoding, and deployment. Within each of these categories, our taxonomy brings together prior works that have been treated as disjoint areas targeting different types of harms (toxic/biased language and misinformation). In addition, we identify their gaps and highlight directions for future research. These include incorporating sociocultural context to produce socially-sensitive interventions, detecting and handling generations with different intents (inadvertent vs. malicious), and going beyond an English, Western/US-centric view to account for the challenges of ethics in *multilingual* language generation.

## 2 Background

Throughout this paper, we use the term *language models* (LMs) to refer to their classic definition as generative models, which predict the next token given the preceding generated context. This paradigm also subsumes conditional LMs that depend on additional inputs via an encoder. We provide more details in Appendix A.

### 2.1 Risks in Language Generation

Before diving into mitigation techniques (§3), we briefly outline potential harms that LMs can cause, following Weidinger et al. (2022)'s taxonomy.

**Discrimination, Toxicity, and Exclusion:** The scope of linguistic diversity in human communication is enormous and is linked to personal, social, and cultural factors (Holmes and Wilson, 2017; Eckert and McConnell-Ginet, 2003; Coates, 2016; Chambers, 1995). As such, language produced in the real world reflects sociocultural stereotypes and presuppositions that LMs can overfit to and amplify (Bar-Tal et al., 2013; Zhao et al., 2017; Sun et al., 2019), leading to several types of harms. (1) *Stereotyping and discrimination* occurs when generated text reinforces discriminatory stereotypes and perpetuates biases against disadvantaged groups, based on factors like gender, race, religion, sexuality, (Bender et al., 2021), and intersectional identities (Crenshaw, 2017). Evidence for this behavior has been substantially corroborated in NLP literature (Blodgett et al., 2020; Nadeem et al., 2021; Nozza et al., 2021; Liang et al., 2021; Field et al., 2021; Lin et al., 2022a, *inter alia*). (2) *Toxicity* describes generated language that is offensive, threatening, violent, or otherwise harmful (Gehman et al., 2020; Rae et al., 2021; Abid et al., 2021). It can range from overtly toxic content, such as violent hate speech, to more subtle, veiled toxicity, such as microaggressions (Breitfeller et al., 2019). (3) *Exclusion* refers to the disparate performance of models across language variations. Models may fail to understand or generate "non-standard" dialects and sociolects, essentially excluding speakers of such variants from their user base (Joshi et al., 2020; Koenecke et al., 2020; Winata et al., 2021).

**Factual Errors, Misinformation, and Disinformation:** LMs are able to generate fluent outputs that users may easily mistake for human-written text (Ippolito et al., 2020), but such utterances may be factually incorrect or misleading (Maynez et al., 2020; Xu, 2020; Lin et al., 2022b; Bickmore et al., 2018; Daws, 2020). This can cause harm inadvertently (via misinformation) or can also be used maliciously (disinformation; Bradshaw and Howard, 2019; Beskow, 2020; Buchanan et al., 2021).

**Privacy Violations:** LMs' vast training corpora often contain sensitive information, and LMs can memorize these details and generate them verbatim when prompted by users, leading to privacy violations (Kim, 2016; Mirshghallah et al., 2020; Brown et al., 2022). LMs have been shown to leak personally identifiable information, such as social security numbers, phone numbers, bank account information (Carlini et al., 2021), and private clinical notes (Lehman et al., 2021); they have even leaked software code and other protected intellectual property (Ippolito et al., 2022). Deploying large LMs can thus pose serious security risks to people whose private information might have found its way into a model's training data.

**Other Underexplored Issues:** Weidinger et al. (2022) discuss other malicious applications, as well as the economical and environmental impacts of LMs. While extremely important, mitigating these risks requires not only technical innovation, but also the development of regulatory practices and policies in an interdisciplinary effort. We focus on algorithmic solutions in this survey, leaving this discussion for future work.

## 3 Taxonomy of Intervention Strategies

The development pipeline of a typical machine learning model involves several critical decisions where risks of harms can arise. Stakeholders have access to different pipeline components and therefore may employ different intervention strategies. For example, while a researcher involved in data curation can intervene before training, an application developer with limited access to a black-box model might only be able to intervene at inference. We present a taxonomy of intervention strategies organized by the stages of a model development lifecycle (Fig. 1), aiming to showcase the tools that can be employed at different stages. We step backward through the pipeline, beginning with application-level interventions employed post-deployment and peeling back the layers through output-level interventions, model interventions, and finally ending at data-level interventions (summarized in Tab. 1).

### 3.1 Application Level Interventions

#### 3.1.1 Harm Detection and Redaction

In order to mitigate harms at the application level, we first need to be able to *detect* problematic, incorrect, and unreliable model outputs (Raji et al., 2020). User-facing applications can employ detectors to intervene before harmful text reaches a user. Such detectors are typically coarse, binary text classifiers, often trained for a single task, such as predicting toxicity (Nobata et al., 2016; Davidson et al., 2017b; Xiang et al., 2021), or the factual accuracy of the outputs (Kryscinski et al., 2020; Goyal and Durrett, 2020; Wang et al., 2020).

Early approaches to building toxicity detectors focused on linear models relying on hand-designed *features* based on lexicons, e.g., hatebase, (Xiang et al., 2012; Dadvar et al., 2012; Burnap and Williams, 2015; Liu and Forss, 2015), $n$-grams, capitalization/punctuation details (Chen et al., 2012; Waseem and Hovy, 2016; Nobata et al., 2016; Xu et al., 2012; Burnap and Williams, 2016). For misinformation detection, features like the presence of new entities or facts in generated document summaries have been employed which can indicate hallucination (Zhao et al., 2020; King et al., 2022).

Linear classifiers, while interpretable, tend to overfit to lexical features, are prone to false positives, and are easy for malicious users to bypass (Kurita et al., 2019). Neural text classifiers, on the other hand, can incorporate contextual information and have been shown to be more robust (Gambäck and Sikdar, 2017; Pitsilis et al., 2018). When built by finetuning pretrained LMs instead of training from scratch, they naturally lead to even better performance (d'Sa et al., 2020; Xiang et al., 2021). Based on these models several toxicity detection tools like Perspective API, OpenAI content filter or ToxiGEN are now publicly available.

To train classifiers for toxicity detection, annotated datasets in several domains have been collected for English (Davidson et al., 2017a; Waseem and Hovy, 2016; Wiegand et al., 2018; Pavlopoulos et al., 2017; Mubarak et al., 2017; Moon et al., 2020), especially to detect overtly toxic text. Human annotation efforts for more subtle toxicities like microaggressions, however, is challenging due to annotators' own biases (Breitfeller et al., 2019). Hence, unsupervised or distantly supervised approaches have been adopted to detect them (Korzeniowski et al., 2019; Field and Tsvetkov, 2020; Sabri et al., 2021). Compared to English, such resources for other languages are severely lacking (Ousidhoum et al., 2019a).

Information-related harms can arise either inadvertently (due to model errors) or deliberately (due to malicious users). Detecting manipulation in the human-written text is an active area of research and those approaches can also be employed for machine-generated text. Prominent research directions include automated fact-checking, propaganda, or fake news detection for which several annotated datasets (Oshikawa et al., 2020; Martino et al., 2020; Zhou and Zafarani, 2020; Guo et al., 2022; Huang et al., 2022) and shared tasks (Thorne et al., 2018; Da San Martino et al., 2019; Feldman et al., 2021) exist. These approaches have also been adopted to assist human fact-checkers (Shaar et al., 2021; Nakov et al., 2021). However, humans are easily fooled by machine-generated fake news (Zellers et al., 2020; Ippolito et al., 2020). An alternate solution is to, not find informational discrepancies, but simply detect and flag whether

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

Table 1: Strategies for mitigating various risks and harms from language models.

the text has been machine-generated (Gehrmann et al., 2019; Dugan et al., 2020; Ippolito et al., 2020; Mitchell et al., 2023), putting the onus to trust the information on the users (Jawahar et al., 2020).

To detect *inadvertent* factual errors, prior works have developed classifiers by training them to detect heuristically introduced synthetic errors in factually correct text (Kryscinski et al., 2020; Goyal and Durrett, 2020), or question-answering errors using targeted QA models (Scialom et al., 2021). Being trained on synthetic data, such detectors typically do not generalize and have low human judgment correlations (Pagnoni et al., 2021).

Relying on the detectors, the most straightforward way a user-facing application can prevent harm is to not display the text at all (*redacting*) or to display it with a warning sign (*flagging*) (Xu et al., 2020). Even when the detectors are imperfect, explicitly flagging problematic outputs is still useful because it signals users to take model outputs with a grain of salt. However, this strategy is not always applicable: for example, in speech-based dialogue agents, "displaying" a warning sign is a nontrivial UX decision, and in auto-complete assistants (such as in Gmail Smart Compose), redacting is not an option and simply warning may not dissuade users from accepting the generated text.

**Challenges:** Predicting whether a text is harmful is often highly contextual and subjective. For toxicity detection, factors like region, political views, and the users' sociocultural background affect whether they perceive the text as toxic (Xenos et al., 2021). Existing datasets are often biased due to their curation process (Dixon et al., 2018; Wiegand et al., 2019; Geva et al., 2019; Sap et al., 2021; Kryscinski et al., 2020) and can have unreliable annotations (Ross et al., 2017; Field and Tsvetkov, 2020; Pagnoni et al., 2021). Further, as with many black-box models, classifiers overfit to spurious artifacts (Gururangan et al., 2018; McCoy et al., 2019; Kumar et al., 2019) and amplify biases in their training data (Zhao et al., 2017; Sun et al., 2019). For instance, toxicity detectors have been shown to disproportionately flag African-American English (AAE) as toxic (Sap et al., 2019). Additionally, such filters might overfit to a subset of small features, with more subtle problematic text evading such filters. Ippolito et al. (2022) show that blocking verbatim training data is insufficient for mitigating privacy concerns in code-generation. We discuss these issues further in §4, highlighting future directions to building finer-grained and explainable approaches for detecting harmful text.

### 3.2 Output Level Interventions

Increasingly, practitioners are building applications using LMs as APIs without explicit knowledge of how the model was trained or what training data was used.[2] Such APIs may vary in how much information developers can see: some allow access to all LM parameters, while black box APIs like GPT3 limit access to model outputs only. Hence, multiple solutions have been proposed for intervening at *model output generation* by editing the outputs with auxiliary models or modifying decoding algorithms.

#### 3.2.1 Post-Factum Editing Model Outputs

Recent studies have explored ways to *edit or revise* model-generated text to remove harmful content. Text editing is a decades-old subfield of NLP that has traditionally focused on fixing errors in machine translation (Chollampatt et al., 2020; Simard et al., 2007; Chatterjee et al., 2020) or grammar in human-written text (Wang et al., 2021c). While many approaches in this area are applicable to post-editing LM outputs, in this survey, we highlight recent work related to *rewriting harmful text*.

The first set of works treats the task of rewriting as a sequence labeling task, where each token in the output sequence is either substituted, deleted, or kept the same (Pryzant et al., 2020; He et al., 2021b). This, however, can be limiting when the entire output needs rewriting. For text-to-text tasks, like translation, summarization, etc. which are trained with parallel data, the same data can be adapted to train an editing model by converting source-target pairs to source-*output*-target triplets using model-generated *outputs* for each source, along with an additional signal indicating errors (obtained using automatic evaluators or human judgment). For more open-ended tasks, prior works explored unsupervised solutions for bias correction (Ma et al., 2020) and semi-supervised methods to correct factual errors (Cao et al., 2020; Lee et al., 2022a; Balachandran et al., 2022). Such methods create synthetic data by inducing errors in clean text and train a model to correct them.

#### 3.2.2 Decoding Methods

Several search and sampling algorithms have been introduced recently to improve the quality of LM-

[2] see https://gpt3demo.com/ for examples

generated text (Graves, 2012; Fan et al., 2018; Holtzman et al., 2020; Meister et al., 2022). In parallel, works on controlling decoding algorithms to promote or demote specific properties in the output text have been developed (Zhang et al., 2022a).

The decoding controls are auxiliary models measuring if the generated text is harmful implemented similarly to the detectors we discussed in §3.1.1, such as toxicity/bias classifiers (Dathathri et al., 2019; Krause et al., 2021; Liu et al., 2021a), factuality metrics (Kryscinski et al., 2020; Goyal and Durrett, 2020). A simple way to use the detectors is *rejection sampling* or *reranking*: for a given input, multiple outputs are generated and then reranked using detector scores to discard dubious outputs (Krishna et al., 2022; King et al., 2022). However, this is often intractable for complex phenomena like factual accuracy of a text or when using multiple controls, since all the generated candidates might be rejected.

To tackle these issues, a class of algorithms that we call *guided-autoregressive decoding* aims to incorporate control by modifying output distributions at every decoding step. One branch of work adopts *logical* controls, where developers directly specify sets of words that should (or not) appear in the output (Lu et al., 2021; Pascual et al., 2021). Wolf et al. (2020) apply this method to zero out the probabilities of offensive terms, King et al. (2022); Lu et al. (2022) improve factual accuracy of generated text by up-weighting generation probabilities of entities present in the source, and Majmudar et al. (2022) apply it for differentially private decoding. A second branch of work composes the LM likelihood with the probabilities from the detectors, to up-weight or down-weight the token probabilities at each decoding step (Yang and Klein, 2021; Liu et al., 2021a; Dathathri et al., 2019; Krause et al., 2021; Schick et al., 2021).

More recent work has also explored ways to induce sentence-level control via *non-autoregressive controlled decoding*. These algorithms incorporate control using Monte Carlo Markov Chain (MCMC) techniques (Hoang et al., 2017; Qin et al., 2020; Mireshghallah et al., 2022), in which a full sequence is initialized and iteratively updated. They have been applied for reducing toxicity (Kumar et al., 2022), and improving fidelity in translation systems (Kumar et al., 2021b). While promising, these techniques suffer from slower decoding speed and need further exploration to be practically used.

**Challenges** Decoding interventions rely on accurate detectors, hence challenges in designing robust detectors (§3.1.1) also impact decoding algorithms. For example, Xu et al. (2021) show that toxicity avoidance algorithms refrain from generating AAE, thereby causing another harm (exclusion) while trying to address the first (toxicity). Also, detecting misinformation and factuality can be extremely hard using simple detectors that do not provide a useful signal to guide the decoding process, so prior works have primarily employed heuristics. Finally, controlled decoding algorithms are double-edged in that controls can be reversed by malicious users to inflict harm—to generate hateful messages, or to do targeted manipulation by copying users' personas. However, this risk should not discourage research in decoding algorithms; rather, research on detecting such malicious uses should be conducted in parallel.

### 3.3 Model Level Interventions

Several recent studies have provided evidence that certain optimization procedures can result in harmful generations downstream (Hall et al., 2022; Taori and Hashimoto, 2022). In this section, we describe approaches that modify LM parameters to prevent such generations by either architecture/training interventions or finetuning/model editing interventions.

#### 3.3.1 Architecture and Training Algorithms

Closely related to applying control at inference time are class-conditioned LMs, which are trained to depend on "control codes" via an additional input (Keskar et al., 2019; Gururangan et al., 2020; Chan et al., 2021). When trained with data annotated for toxicity or bias, these LMs can be prompted to avoid those outputs. Another recently popularized paradigm in LM training is *instruction-based learning*, where in addition to the objective to predict the next token, models are also trained to solve NLP tasks with instructions written in natural language (Wei et al., 2022a; Sanh et al., 2022). Providing explicit instructions to not generate harmful text has shown some promise (Ouyang et al., 2022; Wei et al., 2022a) and is an interesting avenue for future work.

In text-to-text tasks like summarization, the goal is to produce text that is factually consistent with the input without hallucinating information. An LM, however, is typically not constrained to predict tokens grounded in verifiable knowledge, which

can lead to misinformation. Thus, several studies explore modifying LM training objectives to *incorporate factual information* using either knowledge bases (KBs) or graphs (Yu et al., 2022): each token prediction is scored not only on its likelihood given context, but also on whether the generation is grounded in facts in the KBs (Wang et al., 2021b).[3]

However, existing KBs are limited in size as manually curating them is an arduous and expensive process. As an alternative, Liu et al. (2022) propose using automatically generated KBs to train LMs. In contrast, Lewis et al. (2020); de Masson d'Autume et al. (2019); Izacard and Grave (2021) use unstructured text as knowledge. Known as *retrieval-augmented LMs*, they are trained with a two-stage approach of first retrieving a document from an unstructured source like Wikipedia and using it as additional context for generation, essentially providing evidence for the LM-generated text. Wang et al. (2021a); Ji et al. (2020) follow a similar approach to embed commonsense knowledge in LMs. These existing solutions have been used to tackle content-related harms like factual consistency in generated text (Huang et al., 2020; Bapna and Firat, 2019; Dinan et al., 2019; Fan et al., 2021) but future work in reducing discrimination and toxicity in LMs may also benefit from KBs that encode social (Chang et al., 2020), cultural, (Hershcovich et al., 2022), and moral norms (Hendrycks et al., 2021; Jiang et al., 2021). Such LMs augmented with external knowledge can also be dynamically updated by modifying the knowledge source at test time with new information (Khandelwal et al., 2020; He et al., 2021a).

While external knowledge helps provides context, models may not rely on them and still hallucinate. To explicitly control for context, recent studies have explored (1) modifying attention mechanisms to specifically capture relationships between entities (Nan et al., 2021; Zhu et al., 2021), (2) improving coreference to mitigate gender bias in translation (Levy et al., 2021), and (3) using text entailment to develop loss functions to improve fidelity (Falke et al., 2019; Li et al., 2018). Some other notable directions in this space involve fact-aware pretraining (Falke et al., 2019; Wan and Bansal, 2022) and structured learning frameworks (Wiseman et al., 2018).

[3]Knowledge-augmented LMs is a rich field where most existing work focuses on masked LMs (Zhu et al., 2022) for solving understanding tasks. Here we highlight papers on generation.

Finally, to reduce privacy risks in LMs that memorize user information without sacrificing model capabilities, most prominent solutions are based on differentially private (DP) learning (Kerrigan et al., 2020; Shi et al., 2021). DP can provide provable guarantees on the privacy-utility trade-off, however, it requires the LMs to be retrained for each private information that needs to be removed and be quite expensive.

#### 3.3.2 Finetuning and Model Editing

Designing and training models from scratch to mitigate harms can incur heavy environmental and resource costs. In contrast, an alternative branch of work has developed methods for *modifying the model parameters* of already-trained LMs, which requires much fewer resources. An elementary way of doing this is *finetuning* (a subset of) an LM's parameters on small, curated datasets that contain a well-balanced proportion of data for various demographics and filtered for nontoxicity (Gururangan et al., 2020; Chan et al., 2021; Liu et al., 2023). Such balanced and filtered data encourage models correct biases learned from skewed and toxic training data, resulting in safer generated text.

*Prompt-tuning* based methods (Wang et al., 2022) have also shown some success where instead of fine-tuning all the parameters, a prompt (using a small set of parameters) is learned without modifying the rest of the model to perform a task. This paradigm uses the generative power of large LMs, while simultaneously nudging the distribution of generated text toward less harmful content. These approaches have successfully been used to reduce toxicity (Gehman et al., 2020) and exclusion (Chronopoulou et al., 2020; Kumar et al., 2021a). However, finetuning or prompt-tuning on a small dataset may lead to overfitting reducing the general purpose utility of LMs.

Finetuning LMs with *reinforcement learning* (RL) has been suggested as a better alternative (Alabdulkarim et al., 2021; Liu et al., 2021b; Ouyang et al., 2022; Stiennon et al., 2020; Lu et al., 2022; Ramamurthy et al., 2022) for training modern LMs. RL models do not require carefully balanced datasets and can instead learn from discrete rewards such as human feedback (Sun et al., 2020; Ouyang et al., 2022) or auxiliary model-based feedback (Perez et al., 2022). It has been shown to reduce toxic text generated by the models (Bai et al., 2022) and to encourage models to generate more factual text (Mao et al., 2020; Stiennon et al., 2020).

Another less-explored but more computationally practical alternative to finetuning is *model surgery or editing*, which identifies a specific set of neurons that contribute to harmful generations. Culling such parameters has been shown to reduce toxicity (Geva et al., 2022). In a similar vein, De Cao et al. (2021); Mitchell et al. (2022); Meng et al. (2022, 2023) systematically edit model parameters to revise facts memorized by the model. De Cao et al. (2021); Mitchell et al. (2022) use auxiliary editor networks to predict updates to model parameters constrained to revise a fact without changing other facts. Alternatively, Meng et al. (2022, 2023) use interpretability techniques to identify parameters associated with memorizing said facts and edit them locally to revise them.

**Challenges** The biggest argument against mitigation techniques involving training LMs from scratch or augmenting them with knowledge is its cost, making these interventions infeasible for most researchers and practitioners. However, even for organizations with access to large computing resources, research on training safer LMs lags behind research on training ever-larger LMs on raw data. We attribute this to the difficulty of curating KBs, as well as the decreased training and inference speed that comes with such modifications. Finetuning, on the other hand, is less costly but may reduce the general utility of the LMs and has not been shown to be useful in reducing information-related harms. Future work may benefit from drawing on continual (Dhingra et al., 2022) and reinforcement learning (Ouyang et al., 2022) techniques for more practical solutions for large models.

### 3.4 Data Level Interventions

Training any machine learning model requires data, so a natural approach to creating fairer, more reliable LMs is carefully creating balanced training sets that are broadly representative of different worldviews. This requires dedicated and expensive efforts in data curation (Hutchinson et al., 2021; Jo and Gebru, 2020; Kammoun et al., 2022) and novel data pipelines (Denton et al., 2020). Existing works tackling this issue devise semi-automated solutions, which we categorize as follows.

#### 3.4.1 Data Filtration

This simple technique involves removing problematic documents from the training corpus. As training sets can be extremely large, sophisticated neural filters can be prohibitively slow to apply. Hence, most work has utilized simple filters, such as the presence of "unwanted" words (Raffel et al., 2020) or the predictions of linear classifiers (Brown et al., 2020). To mitigate privacy violations, Henderson et al. (2022) construct clean training data by filtering private information and Kandpal et al. (2022); Lee et al. (2022b) filter duplicate training data.

Due to their simplistic setup, these approaches admit many false negatives (failing to detect documents with subtle toxicity) and false positives (erroneously flagging documents that discuss sensitive topics and use hateful speech as examples; additionally, removing data from different dialects like AAE), unintentionally exacerbating risks of marginalization and exclusion (Dodge et al., 2021)). Alternatively, Ngo et al. (2021) train an LM on raw data, then feed the LM manually-curated toxic prompts and filter out documents to which the LM assigns high probability, and then retrain the LM on the filtered corpus.

#### 3.4.2 Data Augmentation

While data filtration aims to remove problematic training samples, data augmentation aims to offset the effect of problematic data by *adding* safer/healthier examples to existing datasets. Mathew et al. (2018) explore adding counterspeech (comments that counter the hateful or harmful speech) to datasets in order to balance out the hate speech already present in web data. Augmentation with synthetically generated data has also been explored for gender bias mitigation in dialogue (Dinan et al., 2020; Liu et al., 2020) and translation models (Stafanovičs et al., 2020).

**Challenges** Since language, identity, and society are tightly intertwined, aggressive data filtering methods risk further imbalancing already imbalanced data. Besides, models trained on filtered data may still degrade when toxic inputs are provided to it. Further, while data augmentation methods have merit, these methods are extremely difficult to large scale. Finally, data interventions are primarily designed to address population-centric risks such as discrimination, toxicity, and, to an extent, exclusion and privacy—but not factuality which is a by-product of training. It is challenging to define (Aly et al., 2022) and detect unsupported facts (Ansar and Goswami, 2021) in the wild, making data interventions insufficient for addressing misinformation and factuality-related harms.

## 4 Discussion and Open Challenges

Though the interventions strategies we discuss achieve some success, many risks of LMs are still not well understood. Below we discuss open problems and avenues for future work to encourage the development of safer LMs.

**Where should one intervene?** Different stakeholders are involved in different model development phases with varying access to resources. As a result, intervention strategies are different depending on the stakeholder. A significant chunk of the responsibility to develop safer LMs falls on researchers and organizations with access to substantial resources who can implement data or modeling interventions. In contrast, practitioners building applications on top of LMs may have access to neither the training data nor the computational resources required to design and train safe LMs. In such cases, flagging and decoding approaches are more practical. In practice, a combination of multiple interventions may be required to both cover a wide array of risks and improve robustness.

**Evolving risks in the ChatGPT era:** LMs are seeing tremendous, rapid growth; larger models are being released every few months (Shoeybi et al., 2019; Brown et al., 2020; Zhang et al., 2022b; Zeng et al., 2023) and deployed in user-facing applications. Many recent LMs like OpenAI's ChatGPT have garnered attention beyond the research community, impacting a range of fields and crossing geographical and language barriers to reach users all over the world (Reuters, 2023; Varghese, 2023; So-hyun, 2023). In such a fast-moving ecosystem, it is ever more essential to proactively study and mitigate LMs' potential harms. Risk mitigation research tends to lag behind model development and is often considered as an afterthought. Though behaviors may emerge unpredictably (Wei et al., 2022b), as we outline in this survey, intervention strategies can and should be applied at different stages of model development to reduce the potential for these influential LMs to cause harm.

**Risks exist in LMs in all languages:** Most research on large LMs, their uses, and their risks is Western-centric and primarily conducted on the English language. However, while a few studies have been conducted on detecting harmful text in non-English datasets (Ousidhoum et al., 2019b; Leite et al., 2020; Burtenshaw and Kestemont, 2021; Bogoradnikova et al., 2021; Costa-jussà et al., 2022, *inter alia*), research on mitigation in non-English settings is lagging (Pamungkas et al., 2021). Further, the definitions of risks themselves change with different context and across cultures. Hence, there is a dire need to develop cross-cultural, cross-lingual analyses as well as mitigation tools.

**Harm detection beyond simple classifiers** Many of the shortcomings of interventions are at their root due to poorly defined risk detection methods. Current detection methods are primarily binary classifiers on various axes like toxicity and factuality, but we recommend researchers and practitioners to move beyond simplistic coarse classifiers and towards more fine-grained (Xiang et al., 2021; Goyal and Durrett, 2020; Da San Martino et al., 2020), interpretable (Koh and Liang, 2017; Han and Tsvetkov, 2020, 2021), and explainable (Pagnoni et al., 2021; Gehrmann et al., 2019) harm detectors to support better harm mitigation strategies (Lipton, 2018; Jacovi et al., 2021).

**Systematic evaluation frameworks for mitigation strategies** Though LM performance is usually systematically evaluated through benchmarks (Wang et al., 2019b,a; BIG-bench collaboration, 2022), practices for evaluating harms in LM-generated text or the effectiveness of mitigation strategies are not. While there is an emerging body of work dedicated to benchmarking LM harms (Rauh et al., 2022), the space of potential harms is huge and intersectional, and existing work only covers a fraction of it. Developing a suite of evaluations or augmenting existing generation benchmarks (Mille et al., 2021) with axes of risk evaluations (Ribeiro et al., 2020) will encourage the development of holistic solutions, bridging discrimination/toxicity and information-related harms—two related directions in which researchers have often developed similar solutions.

## 5 Conclusion

We present a survey of practical methods and techniques for addressing the societal harms and safety risks of language generation models. Our structured taxonomy covers a wide variety of interventions at different stages of the model development pipeline to mitigate harms. This work bridges multiple strands of research and presents an actionable overview on methods for preventing harms from language generation models.

## Limitations

The goal of this survey was to present current research on analyzing and mitigating harms of language generation. There are multiple documented and anticipated harms that these models perpetuate, and it is not feasible to address intervention strategies for each of them. We aimed to generalize multiple proposed solutions and present them in a structured form, considering a few popularly studied harms as case studies. Inevitably, certain harms and their mitigation strategies might not have been considered for this survey.

Current research in this field is nascent but fast-moving. While this survey enlists techniques and approaches that are popular now, there is a potential for them to be replaced with newer research. We anticipate that this survey may need to be updated or even redone to incorporate new research.

## Ethics Statement

In this survey, we present and discuss various risk analyses and intervention strategies to prevent societal harms from LMs. We also comment on common themes across approaches for detecting and resolving population-centric harms (such as toxicity and discrimination) and misinformation-related harms, and we recommend future work combining them. First, many datasets and resources we discuss may contain biases, and using them in downstream applications can lead to risks as we have outlined. Second, many techniques we discuss have limitations or are known to exacerbate other kinds of harms (Xia et al., 2020), and thus, applying them to newer problems may lead to unseen issues. Finally, the interventions we identify to raise general awareness have the potential for misuse: a malicious user can further imbalance the data to train even more harmful models, use the models and decoding algorithms to generate fake news, and target marginalized populations. This, however, should not discourage the development of mitigation strategies; rather, more work should be done to detect and ban malicious users. This requires not only technological solutions in NLP, but also in social science, social network analysis, and public policy.

## Acknowledgements

We gratefully acknowledge support from NSF CAREER Grant No. IIS2142739, NSF grants No. IIS2125201, IIS2040926, IIS2203097, Workhuman, and an Alfred P. Sloan Foundation Fellowship. S.K. gratefully acknowledges support from a Google PhD Fellowship.

## References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. *Persistent Anti-Muslim Bias in Large Language Models*, page 298–306. Association for Computing Machinery, New York, NY, USA.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2022. HTLM: Hyper-text pre-training and prompting of language models. In *International Conference on Learning Representations*.

Amal Alabdulkarim, Winston Wai-Tai Li, Lara J. Martin, and Mark O. Riedl. 2021. Goal-directed story generation: Augmenting generative language models with reinforcement learning. *ArXiv*, abs/2112.08593.

Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors. 2022. *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*. Association for Computational Linguistics, Dublin, Ireland.

Wazib Ansar and Saptarsi Goswami. 2021. Combating the menace: A survey on characterization and detection of fake news from a data science perspective. *International Journal of Information Management Data Insights*, 1(2):100052.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ankur Bapna and Orhan Firat. 2019. Non-parametric adaptation for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1921–1931, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Bar-Tal, Carl F Graumann, Arie W Kruglanski, and Wolfgang Stroebe. 2013. *Stereotyping and prejudice: Changing conceptions*. Springer Science & Business Media.

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*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

David Beskow. 2020. Finding and Characterizing Information Warfare Campaigns.

Timothy Bickmore, Ha Trinh, Stefan Olafsson, Teresa O'Leary, Reza Asadi, Nathaniel Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant. *J Med Internet Res*, 20:e11510.

BIG-bench collaboration. 2022. Beyond the imitation game: Measuring and extrapolating the capabilities of language models.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Darya Bogoradnikova, Olesia Makhnytkina, Anton Matveev, Anastasia Zakharova, and Artem Akulov. 2021. Multilingual sentiment analysis and toxicity detection for text messages in russian. In *2021 29th Conference of Open Innovations Association (FRUCT)*, pages 55–64.

Samantha Bradshaw and Philip N Howard. 2019. The global disinformation order: 2019 global inventory of organised social media manipulation.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Hannah Brown, Katherine Lee, FatemehSadat Mireshghallah, R. Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. 2021. Truth, lies, and automation.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242.

Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15.

Ben Burtenshaw and Mike Kestemont. 2021. A Dutch dataset for cross-lingual multilabel toxicity detection. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 75–79, Online (Virtual Mode). INCOMA Ltd.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

J. K. Chambers. 1995. *Sociolinguistic theory: linguistic variation and its social significance*. Oxford.

Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2021. Cocon: A self-supervised approach for controlled text generation. In *Proc. ICLR*.

Ting-Yun Chang, Yang Liu, Karthik Gopalakrishnan, Behnam Hedayatnia, Pei Zhou, and Dilek Hakkani-Tur. 2020. Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks. In *Proceedings of Deep Learning Inside Out*

*(DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 74–79, Online. Association for Computational Linguistics.

Open AI's Assistant ChatGPT and Andrew M. Perlman. 2022. The implications of openai's assistant for legal services and society. *SSRN Electronic Journal*.

Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020. Findings of the WMT 2020 shared task on automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 646–659, Online. Association for Computational Linguistics.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Shamil Chollampatt, Raymond Hendy Susanto, Liling Tan, and Ewa Szymanska. 2020. Can automatic post-editing improve nmt? *arXiv preprint arXiv:2009.14395*.

Dennis Chong and James N Druckman. 2007. Framing theory. *Annual review of political science*, 10(1):103–126.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2020. Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711, Online. Association for Computational Linguistics.

Jennifer Coates. 2016. *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*. Routledge.

Marta R Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Javier Ferrando, and Carlos Escolano. 2022. Toxicity in multilingual machine translation at scale. *arXiv preprint arXiv:2210.03070*.

Kimberlé W. Crenshaw. 2017. *On Intersectionality: Essential Writings*. The New Press, New York, NY.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170, Hong Kong, China. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Maral Dadvar, FMG de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *ICLR*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017a. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017b. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Ryan Daws. 2020. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Ashwin Geet d'Sa, Irina Illina, and Dominique Fohr. 2020. Bert and fasttext embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*, pages 1–5. IEEE.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 2020. RoFT: A tool for evaluating human detection of machine-generated text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 189–196, Online. Association for Computational Linguistics.

Penelope Eckert and Sally McConnell-Ginet. 2003. *Language and Gender*. Cambridge University Press.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *ACL (1)*, pages 2214–2220.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2021. Augmenting Transformers with KNN-Based Composite Memory for Dialog. *Transactions of the Association for Computational Linguistics*, 9:82–99.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Anna Feldman, Giovanni Da San Martino, Chris Leberknight, and Preslav Nakov. 2021. Proceedings of the fourth workshop on nlp for internet freedom: Censorship, disinformation, and propaganda. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. 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)*, pages 1905–1925, Online. Association for Computational Linguistics.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

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*, pages 3356–3369, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Ke Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *ArXiv*, abs/2203.14680.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language under-

standing datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166.

Nancy Gleason. 2022. Chatgpt and the rise of ai writers: how should higher education respond?

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Melissa R.H. Hall, Laurens van der Maaten, Laura Gustafson, and Aaron B. Adcock. 2022. A systematic study of bias amplification. *arXiv preprint arXiv:2301.11305*.

Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying toxic speech detectors against veiled toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7732–7739, Online. Association for Computational Linguistics.

Xiaochuang Han and Yulia Tsvetkov. 2021. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4398–4409, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021a. Efficient nearest neighbor language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5703–5714, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zexue He, Bodhisattwa Prasad Majumder, and Julian McAuley. 2021b. Detect and perturb: Neutral rewriting of biased and sensitive text via gradient-based decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4173–4181, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *NeurIPS*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. 2017. Towards decoding as continuous optimisation in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 146–156, Copenhagen, Denmark. Association for Computational Linguistics.

Janet Holmes and Nick Wilson. 2017. *An introduction to sociolinguistics*. Routledge.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Nabil Hossain, Marjan Ghazvininejad, and Luke Zettlemoyer. 2020. Simple and effective retrieve-edit-rerank text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2532–2538, Online. Association for Computational Linguistics.

Kung-Hsiang Huang, Preslav Nakov, Yejin Choi, and Heng Ji. 2022. Faking fake news for real fake news detection: Propaganda-loaded training data generation. *ArXiv*, abs/2203.05386.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107, Online. Association for Computational Linguistics.

Elle Hunt. 2016. Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA. Association for Computing Machinery.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini. 2022. Preventing verbatim memorization in language models gives a false sense of privacy. *ArXiv*, abs/2210.17546.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. *Proc. FAccT*.

Heesoo Jang. 2021. A South Korean chatbot shows just how sloppy tech companies can be with user data.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2022. Knowledge unlearning for mitigating privacy risks in language models. *ArXiv*, abs/2210.01504.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736, Online. Association for Computational Linguistics.

Liwei Jiang, Jena D. Hwang, Chandrasekhar Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards machine ethics and norms. *ArXiv*, abs/2110.07574.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA. Association for Computing Machinery.

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*, pages 6282–6293, Online. Association for Computational Linguistics.

Amina Kammoun, Rim Slama, Hedi Tabia, Tarek Ouni, and Mohmed Abid. 2022. Generative adversarial networks for face generation: A survey. *ACM Computing Surveys*.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. *ICML*.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training. In *Proceedings of the Second Workshop on Privacy in NLP*, pages 39–45, Online. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *ACM Computing*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Dongwoo Kim. 2016. Chatbot gone awry starts conversations about ai ethics in south korea.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S Weld, Iz Beltagy, and Doug Downey. 2022. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search. *arXiv preprint arXiv:2203.08436*.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Renard Korzeniowski, Rafal Rolczynski, Przemyslaw Sadownik, Tomasz Korbak, and Marcin Możejko. 2019. Exploiting unsupervised pre-training and automated feature engineering for low-resource hate speech detection in polish. *ArXiv*, abs/1906.09325.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kalpesh Krishna, Ya yin Chang, John Wieting, and Mohit Iyyer. 2022. Rankgen: Improving text generation with large ranking models. *EMNLP*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021a. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 110–121, Online. Association for Computational Linguistics.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021b. Controlled text generation as continuous optimization with multiple constraints. In *Proc. NeurIPS*.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. 2022. Gradient-based constrained sampling from language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2251–2277.

Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4153–4163, Hong Kong, China. Association for Computational Linguistics.

Keita Kurita, Anna Belova, and Antonios Anastasopoulos. 2019. Towards robust toxic content classification. *arXiv preprint arXiv:1912.06872*.

Hwanhee Lee, Cheoneum Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Juae Kim, and Kyomin Jung. 2022a. Factual error correction for abstractive summaries using entity retrieval. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 439–444, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022b. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. 2021. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959, Online. Association for Computational Linguistics.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive

sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori B. Hashimoto. 2022. Large language models can be strong differentially private learners. *ICLR*.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Inna Wanyin Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina C. H. Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022a. Gendered mental health stigma in masked language models. *EMNLP*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Ves Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. *ArXiv*, abs/2112.10668.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021a. DExperts: Decoding-time controlled text generation with experts and anti-experts. 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)*, pages 6691–6706, Online. Association for Computational Linguistics.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Qi Liu, Dani Yogatama, and Phil Blunsom. 2022. Relational Memory-Augmented Language Models. *Transactions of the Association for Computational Linguistics*, 10:555–572.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021b. Mitigating political bias in language models through reinforced calibration. In *AAAI*.

Shuhua Liu and Thomas Forss. 2015. New classification models for detecting hate and violence web content. In *2015 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3K)*, volume 1, pages 487–495. IEEE.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. QUARK: Controllable text generation with reinforced unlearning. In *Advances in Neural Information Processing Systems*.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online. Association for Computational Linguistics.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Jimit Majmudar, Christophe Dupuy, Charith S. Peris, Sami Smaili, Rahul Gupta, and Richard S. Zemel. 2022. Differentially private decoding in large language models. *ArXiv*, abs/2205.13621.

Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. 2020. Constrained abstractive summarization: Preserving factual consistency with constrained generation. *ArXiv*, abs/2010.12723.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Binny Mathew, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2018. Thou shalt not hate: Countering online hate speech. In *International Conference on Web and Social Media*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proc. ACL*.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Typical decoding for natural language generation. *ArXiv*, abs/2202.00666.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *International Conference on Learning Representations*.

Simon Mille, Kaustubh Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic construction of evaluation suites for natural language generation datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generationusing energy language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 401–415, Dublin, Ireland. Association for Computational Linguistics.

Fatemehsadat Mirshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh. 2020. Privacy in deep learning: A survey. *ArXiv*, abs/2004.12254.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the first workshop on abusive language online*, pages 52–56.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language 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)*, pages 5356–5371, Online. Association for Computational Linguistics.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barr'on-Cedeno, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *International Joint Conference on Artificial Intelligence*.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. 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)*, pages 6881–6894, Online. Association for Computational Linguistics.

Helen Ngo, Cooper Raterink, João GM Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. 2021. Mitigating harm in language models with conditional-likelihood filtration. *arXiv preprint arXiv:2108.07790*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019a. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

*9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019b. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *NeurIPS*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing*, pages 1–27.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sajan Patel and Kyle Lam. 2023. Chatgpt: the future of discharge summaries? *The Lancet. Digital health*.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35, Vancouver, BC, Canada. Association for Computational Linguistics.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nathan McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Conference on Empirical Methods in Natural Language Processing*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 794–805, Online. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021.

Scaling language models: Methods, analysis & insights from training gopher. arXiv.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. arXiv.

Maribeth Rauh, John F J Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, William Isaac, and Lisa Anne Hendricks. 2022. Characteristics of harmful text: Towards rigorous benchmarking of language models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Reuters. 2023. Chatgpt sets record for fastest-growing user base.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Nazanin Sabri, Valerio Basile, Tommaso Caselli, et al. 2021. Leveraging bias in pre-trained word embeddings for unsupervised microaggression detection. In *Italian Conference on Computational Linguistics 2021: CLiC-it 2021*. CEUR Workshop Proceedings (CEUR-WS. org).

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and A Noah Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Computing Research Repository*, arXiv:2103.00453.

Katherine Schulten. 2023. How should schools respond to chatgpt?

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. In *Conference on Empirical Methods in Natural Language Processing*.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2021. Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. In *Conference on Empirical Methods in Natural Language Processing*.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. 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)*, pages 4275–4293, Online. Association for Computational Linguistics.

Weiyan Shi, Aiqi Cui, Evan Li, R. Jia, and Zhou Yu. 2021. Selective differential privacy for language modeling. In *North American Chapter of the Association for Computational Linguistics*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. arXiv.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *ArXiv*, abs/1909.08053.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing. In *WMT@ACL*.

Kim So-hyun. 2023. [the korean dilemma] what chatgpt means for korea.

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*.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan J. Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *NeurIPS*, abs/2009.01325.

Chris Stokel-Walker. 2023. Chatgpt listed as author on research papers: many scientists disapprove.

Fan-Keng Sun, Cheng-Hao Ho, and Hung yi Lee. 2020. Lamol: Language modeling for lifelong language learning. In *ICLR*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Rohan Taori and Tatsunori Hashimoto. 2022. Data feedback loops: Model-driven amplification of dataset biases. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ranjana Mary Varghese. 2023. Will chatgpt shake up higher education in india?

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

James Vincent. 2022. YouTuber trains AI bot on 4chan's pile o' bile with entirely predictable results.

David Wan and Mohit Bansal. 2022. Factpegasus: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *NAACL 2022*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Boxin Wang, Wei Ping, Chaowei Xiao, Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Bo Li, Anima Anandkumar, and Bryan Catanzaro. 2022. Exploring the limits of domain-adaptive training for detoxifying large-scale language models. *NeurIPS*.

Han Wang, Yang Liu, Chenguang Zhu, Linjun Shou, Ming Gong, Yichong Xu, and Michael Zeng. 2021a. Retrieval enhanced model for commonsense generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3056–3062, Online. Association for Computational Linguistics.

Luyu Wang, Yujia Li, Ozlem Aslan, and Oriol Vinyals. 2021b. WikiGraphs: A Wikipedia text - knowledge graph paired dataset. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 67–82, Mexico City, Mexico. Association for Computational Linguistics.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021c. A comprehensive survey of grammatical error correction. *ACM Trans. Intell. Syst. Technol.*, 12(5).

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. *TMLR*.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 602–608.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187, Brussels, Belgium. Association for Computational Linguistics.

Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on Microsoft's Tay "experiment," and wider implications. *The ORBIT Journal*, 1(2):1–12.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Alexandros Xenos, John Pavlopoulos, and Ion Androutsopoulos. 2021. Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 140–145, Online. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984.

Tong Xiang, Sean MacAvaney, Eugene Yang, and Nazli Goharian. 2021. ToxCCIn: Toxic content classification with interpretability. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 1–12, Online. Association for Computational Linguistics.

Adrian Yijie Xu. 2020. Generating fake news with openai's language models.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proc. NAACL*.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Comput. Surv.* Just Accepted.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending against neural fake news. *Neurips*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, P. Zhang, Yuxiao Dong, and Jie Tang. 2023. Glm-130b: An open bilingual pre-trained model. *ICLR*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022a. A survey of controllable text generation using transformer-based pre-trained language models. arXiv.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020a. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020b. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. Opt: Open pre-trained transformer language models. arXiv.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proc. of EMNLP*, pages 2979–2989.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

Chenguang Zhu, Yichong Xu, Xiang Ren, Bill Yuchen Lin, Meng Jiang, and Wenhao Yu. 2022. Knowledge-augmented methods for natural language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 12–20, Dublin, Ireland. Association for Computational Linguistics.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *ArXiv*, abs/2301.12867.

## A Background: More Details

Since we focus on language generation, we use the term *language models* (LMs) to refer to their classic definition as generative models (or decoders), which predict the next token given the preceding generated context. For the purposes of this survey, this paradigm also subsumes conditional (or sequence-to-sequence) LMs conditioned on inputs from different modalities such as text, image, or speech via an encoder. [4] Unless otherwise specified, we assume that (1) the LM decoder is parameterized by a transformer architecture (Vaswani et al., 2017), and (2) the LM is first pretrained on a large amount of text (ranging from 100-billions to trillions of tokens), which, together with their large number of parameters, have earned such models the name large language models.[5]. After pretraining, LMs are either used in a zero- or few-shot manner (Brown et al., 2020), or modified for specific tasks via finetuning all or some of their parameters (Liu et al., 2023).

The generation tasks this survey focuses on can be broadly categorized as either (1) transformation tasks, where a given input is transformed into a textual output such as machine translation, abstractive summarization, data-to-text generation, and stylistic re-writing, among others (Prabhumoye et al., 2018; Raffel et al., 2020; Zhang et al., 2020b; Aghajanyan et al., 2022), (2) or open-ended tasks such as dialogue generation, prompt-based autocompletion, story generation, and more (Adiwardana et al., 2020; Guan et al., 2020).

[4] While many different strategies to (pre-)train encoder LMs have been introduced in the literature (Devlin et al., 2018; Peters et al., 2018), they are generally not conducive to generating text and are out of scope in this survey.

[5] While some of the studies we will discuss do not rely on pretraining, we highlight it here since it is one of the primary drivers of recent advances in language generation (and its associated risks)