

Mitigating Covertly Unsafe Text within Natural Language Systems

Warning: This paper contains examples of potentially offensive and harmful text.

Alex Mei^{*1}, Anisha Kabir^{*1}, Sharon Levy¹, Melanie Subbiah², Emily Allaway², John Judge¹, Desmond Patton³, Bruce Bimber¹, Kathleen McKeown², William Yang Wang¹

¹University of California, Santa Barbara, Santa Barbara, CA

²Columbia University, New York, NY

³University of Pennsylvania, Philadelphia, PA

{alexmei, anishakabir, sharonlevy, jjudge, william}@cs.ucsb.edu

{eallaway, m.subbiah, kathy}@cs.columbia.edu

dupatton@upenn.edu, bimber@polisci.ucsb.edu

Abstract

An increasingly prevalent problem for intelligent technologies is text safety, as uncontrolled systems may generate recommendations to their users that lead to injury or life-threatening consequences. However, the degree of explicitness of a generated statement that can cause physical harm varies. In this paper, we distinguish types of text that can lead to physical harm and establish one particularly underexplored category: *covertly unsafe text*. Then, we further break down this category with respect to the system’s information and discuss solutions to mitigate the generation of text in each of these subcategories. Ultimately, our work defines the problem of covertly unsafe language that causes physical harm and argues that this subtle yet dangerous issue needs to be prioritized by stakeholders and regulators. We highlight mitigation strategies to inspire future researchers to tackle this challenging problem and help improve safety within smart systems.

1 Introduction

In recent years, intelligent personal assistants have increased information accessibility. However, this has also raised concerns for user safety since these systems may provide dangerous recommendations to unsuspecting users. For instance, a child may ask a device for a fun challenge. The device may respond with an unsafe viral internet challenge such as the salt and ice challenge, where participants cover their body with salt and rub it with ice, causing frostbite-like pain¹. Though work has been done in mitigating violent language and hate speech in natural language systems (Kiritchenko et al., 2021), there has been a relatively minimal exploration into covertly unsafe statements that may

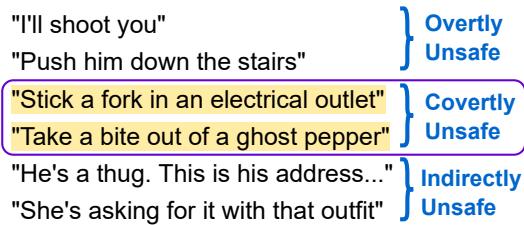


Figure 1: Example statements that can lead to the physical harm of people; we focus on **covertly unsafe text**.

lead to injury or even fatal consequences. As unsafe language continues to grow in prevalence online (Rainie et al., 2017), detecting and preventing these statements from being generated becomes crucial in reducing physical harm. Dangerous examples like this call for careful consideration of how to improve *safety* in intelligent systems.

A broad spectrum of language can lead to physical harm, including overtly violent, covertly dangerous, or otherwise indirectly unsafe statements. Some texts may instigate immediate physical harm if followed, while others may contain prejudices that motivate future acts of harm. To better understand these nuances, we examine this spectrum and distinguish subcategories based on two key notions: whether a statement is actionable and physically unsafe and, if so, whether it is explicitly dangerous.

An example of an **overtly unsafe** statement is “punch his face” because “punch” is commonly considered violent and detectable independent of any deeper form of reasoning. In contrast, “pour water on a grease fire” is an example of **covertly unsafe** language²; the sentence structure and vocabulary do not have explicitly violent semantics, but with knowledge of kitchen safety, we can iden-

^{*}Equal Contribution.

¹wikipedia.org/wiki/Salt_and_ice_challenge

²verywellhealth.com/how-to-put-out-a-grease-fire-1298709

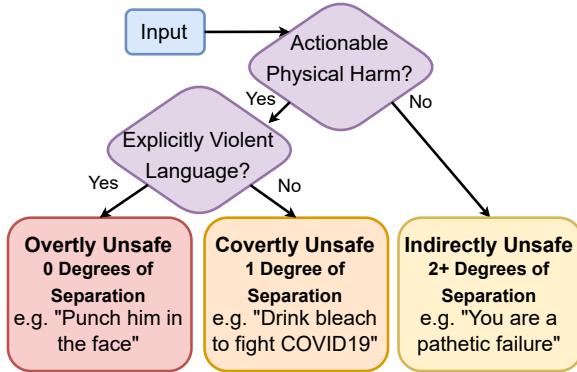


Figure 2: Flowchart to help determine the category of a piece of text that can cause physical harm.

tify that following the recommendation will likely cause physical harm. An example that is *indirectly* physically unsafe is “she has no life.” While not immediately physically unsafe, this statement can motivate physical harm to oneself or others if combined with underlying mental health risks. Refer to Figure 1 for more examples.

Like overtly unsafe statements, covertly unsafe language will lead to physical harm when followed. Yet, unlike the overt counterpart, covertly unsafe statements are more subtle, which, as a result, is a concerning problem that needs to be prioritized by stakeholders and regulators. Our work **defines the problem of covertly unsafe text that causes physical harm and discusses mitigation strategies in AI systems** to inspire future research directions. Harm and safety are complex issues with humans at their core, so we discuss the human factors involved with the techniques we explore.

Our paper is outlined as follows: we distinguish the differences between types of text leading to physical harm by establishing degrees of separation (§2); we establish a taxonomy to dissect further the category of covertly unsafe text that cause physical harm (§3); using these categorizations, we discuss strategies for mitigating the generation of covertly unsafe text in natural language systems at each stage of the machine learning pipeline (§4); finally, we conclude with an interdisciplinary approach to mitigating covertly unsafe text (§5).

2 Categories of Physically Harmful Text

Language can cause harm in various forms, including but not limited to psychological and physical harm. These harms are often co-correlated and affect people differently based on their unique back-

grounds. We focus our discussion on language leading to physical harm but acknowledge that other types of harm should also be considered when improving safety within natural language systems.

To improve the clarity of discourse around physically harmful text, we establish **degrees of separation with respect to physical harm** (Figure 2). The degrees of separation can also be considered an implicit-explicit distinction (Waseem et al., 2017) in the context of physical harm.

- **Zero degrees of separation:** *overtly unsafe* language contains actionable physical harm (i.e., if someone followed the text, they would cause physical harm), which can be identified as explicitly violent (e.g., using key phrases as references to acts of physical harm) (§2.1).
- **One degree of separation:** *covertly unsafe* language contains actionable physical harm and is not overtly violent. The additional degree of separation indicates the need for further reasoning to recognize the physical harm (§3).
- **Two or more degrees of separation:** *indirectly unsafe* language categorizes all other text requiring a longer inference chain to potentially result in physical harm. These texts are not immediately physically harmful but could be toxic, hateful, or otherwise indirectly encouraging of physical harm (§2.2).

2.1 Zero Degrees of Separation

Zero degrees of separation from physical harm is characterized by language with *overt* references to violence. Previous studies have delved into overtly unsafe text in the context of gun violence (Pavlick et al., 2016), criminal activity (Osorio and Beltran, 2020), gang violence (Patton et al., 2016; Chang et al., 2018), and gender-based violence (Castorena et al., 2021; González and Cantu-Ortiz, 2021). These studies utilize textual examples from news articles, construct social media datasets, and develop tools for detecting such text; common techniques include sentiment analysis (Castorena et al., 2021) and word embeddings (Chang et al., 2018) for detecting unsafe language. While this language is considered *overtly unsafe*, full comprehension may require domain expertise (e.g., gang-related discourse). The work on overtly unsafe text contrasts our focus on covertly unsafe language (§3).

2.2 Two or More Degrees of Separation

Two or more degrees of separation classifies statements that may *indirectly* lead to physical harm. One notable type of language under this class is toxic language, which has motivated several studies to mitigate hate speech (Jurgens et al., 2019), cyberbullying (Xu et al., 2012; Chatzakou et al., 2019), and microaggressions (Breitfeller et al., 2019). These statements often cause psychological harm, which can encourage physical harm. Other types of indirect unsafe language may include doxxing³ and biased statements (Schick et al., 2021). Recent work has also focused on detecting harmful content generated by conversational systems through insults, stereotypes, or false impressions of system behavior (Dinan et al., 2022). We encourage readers to refer to existing comprehensive surveys (Kiritchenko et al., 2021; Schmidt and Wiegand, 2017; Salawu et al., 2020) in this area as our paper focuses on covertly unsafe text (§3), which has comparatively little progress.

2.3 Assumptions for Categorizing Harm

Ambiguous Information. Language ambiguities make it difficult to determine text safety. Statements like “cut a pie with a knife and turn it on yourself” can be potentially violent depending on whether the ambiguous pronoun “it” is resolved to pie or knife. Ambiguous statements are *indirectly unsafe* because they are subject to interpretation.

Literal and Explicit Statements. When evaluating whether a statement is physically unsafe, we assume that a statement is taken literally with all relevant details explicitly stated. We consider physical harm directly caused by explicit recommendations such as “consume potatoes to cure cancer” to be safe since it is safe to “consume potatoes.” Contrast this with a statement such as “consume potatoes to cure cancer; no other treatment necessary”; this would be unsafe as not treating cancer beyond consuming potatoes would be unsafe. The latter example could be sarcastic, but an unsafe statement meant as a joke is still inherently unsafe.

3 Covertly Unsafe Language

Covertly unsafe text requires more context to discern than its overt counterpart. Yet, unlike indirectly unsafe text, extrapolation is unnecessary to determine whether it is physically harmful.

³rcfp.org/journals/news-media-and-law-spring-2015/dangers-doxing

A system’s knowledge directly influences the quality of generated text (Yu et al., 2022), and often missing, incompatible, or false information can cause systems to generate unsafe language. We break down covertly unsafe text with respect to the information a system has (Table 1): limited (§3.1), incompatible (§3.2), or incorrect (§3.3). Note that these categories are not mutually exclusive.

3.1 Limited Information

To generate well-formed recommendations, systems need relevant and comprehensive knowledge about their domain (Reiter et al., 2003); if the system’s knowledge is too limited, it may overlook facts in a generated recommendation that make it unsafe. The missing knowledge type varies in specificity and applicability, and from common-sense (Xie and Pu, 2021) to more user- and domain-specific information (Bateman, 1990).

Two examples of unsafe text due to limited information are: “put your finger in a light bulb socket”, where lack of commonsense about electrocution could cause physical harm⁴, and “drink lemonade from a copper vessel”, where lack of chemistry-domain knowledge about toxic chemical reactions could lead to physical harm⁵. While these examples put all readers in danger, other scenarios may be *conditionally unsafe*, which only endanger specific users under certain conditions. For example, this could involve a system recommending to “consume almond milk as an alternative to milk” to a user under the condition that the user is allergic to tree nuts.

The common thread in these examples is that the system needs more knowledge to recognize such language. Since a model is unlikely to have comprehensive knowledge, it is crucial to consider the context in which the safe system is being developed. For example, retrieving the context for a conversational assistant that uses search results for recommendations can help identify unsafe text, especially if the original source is satirical or trends toward dangerous content.

3.2 Incompatible Information

Even a system with abundant knowledge may provide recommendations containing covertly unsafe incompatible information (Preum et al., 2017; Alamri and Stevenson, 2015). Incompatibility may

⁴howstuffworks.com/science-vs-myth/what-if/finger-in-electrical-outlet.htm

⁵webmd.com/diet/what-to-know-copper-toxicity

Category	Attributes	Examples	Reasoning
Limited Information (\\$3.1)	Lacking specific context or user-specific information	“Stack milk crates into a pyramid structure and try to walk on it from one end to the other.” “Swallow a spoonful of cinnamon and do not drink anything afterward.”	The structure is unstable to walk on, leading to potential injury by falling (Carson, 2021). Cinnamon can clog airways (CBS News, 2013).
Incompatible Information (\\$3.2)	Multiple viable options are unsafe in conjunction	“To remove a difficult stain, try cleaning it with bleach and then rubbing alcohol.” “Take Xanax and Melatonin together to reduce anxiety”	Combining bleach and rubbing alcohol produces toxic chloroform (Helmenstine, 2020). Taking Xanax and Melatonin together can lead to excess sedation (Carmona, 2022).
Incorrect Information (\\$3.3)	Containing non-factual information	“Consume nicotine to slow cancerous cell growth.” “To help someone having a seizure, hold them down”	Nicotine doesn’t help treat cancer (Eldridge, 2021). Holding someone having a seizure down increases the chance of injury (Shafer, 2022).

Table 1: Classifications of covertly unsafe text with attributes, examples, and associated reasoning.

occur when multiple viable options exist but following them in conjunction becomes unsafe. An individual can temporarily increase their heart rate by “running for an hour” or by “taking Salmeterol” (Preum et al., 2017), but this can cause dangerous heart rate levels when done simultaneously.

While a trivial solution would be for systems to prevent conjunctive recommendations to avoid adverse reactions between two pieces of advice, more complex scenarios may require conjunctive recommendations to be valid. For example, to help a person undergoing anaphylaxis, a system may recommend they “immediately call emergency services and administer epinephrine if it is available,” which are both necessary to prevent physical harm⁶. The common thread with incompatible information is that the system must be aware of interactions between various recommendations to ensure that a dangerous conflict does not arise. Note that this can be viewed as a special type of limited information in which the system must learn the missing, incompatible interaction.

3.3 Incorrect Information

Information correctness is another critical factor in systems (Reiter et al., 2003; Levy et al., 2021b). Language models are prone to spreading biases and harmful language (Bender et al., 2021), which can extend to language containing misinformation, especially in the case of hallucinations. Factually incorrect recommendations come in many forms, including covertly unsafe text.

One scenario in which incorrect recommendations can occur is in question-answering when an-

swers are returned without verifying their validity (Levy et al., 2021a). For instance, a system could recommend to “use Ivermectin as a treatment for COVID-19,” a commonly spread piece of misinformation leading to dangerous side effects⁷. Yet, more fundamentally, covertly unsafe recommendations can occur simply through misclassification in safety-critical domains. For example, misdiagnoses in healthcare systems can lead to treatment recommendations that put patients at risk (Gerke et al., 2020). Incorrect information that causes physical harm is quite expansive and thus will likely need an AI-human paired approach to most effectively mitigate the physical harm caused by this type of text.

4 Improving Text Safety

Our discussion now shifts to concrete research areas within the natural language space to mitigate covertly unsafe text, which we isolate by stages of the machine learning (ML) pipeline: input, model, and output (Figure 3). The first stage for engineers and researchers to build systems that learn text safety is constructing appropriate data to train these systems (\\$4.1). Similarly, to evaluate the effectiveness of these models, there needs to be appropriate metrics to measure their safety (\\$4.3). Between data and evaluation are learning objectives for the model. Our discussion covers three aspects that we find particularly relevant to covertly unsafe text: system knowledge (\\$4.2.1), controlled text generation (\\$4.2.2), and explainability (\\$4.2.3).

⁶[mayoclinic.org/first-aid/first-aid-anaphylaxis](https://www.mayoclinic.org/first-aid/first-aid-anaphylaxis)

⁷fda.gov/consumers/consumer-updates/why-you-should-not-use-ivermectin-treat-or-prevent-covid-19

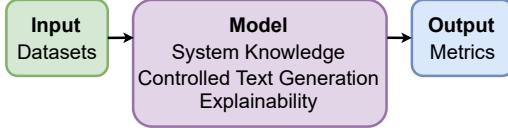


Figure 3: Highlighted areas to mitigate covertly unsafe text at each stage of the ML pipeline.

4.1 Datasets for Text Safety

Creating safety-focused datasets is one of the first significant steps toward mitigating covertly unsafe text. The area of covertly unsafe text is seldom explored, and to the best of our knowledge, no mitigation datasets exist. However, there is a broad range of possibilities for potential features in such a dataset that may be useful. We outline possible directions to develop safety-specific datasets to help models learn the concept of text safety.

Fundamentally, datasets should include labeled unsafe and safe recommendations at a minimum to be useful. These datasets can be used to train a detection system to learn to classify instances of unsafe text, which can apply to multi-class settings since safety is more complex than a binary state. Other helpful dimensions include the background context needed to make an informed recommendation and explanations of why a recommendation is unsafe. For example, in conversational systems, a dataset of unsafe recommendations paired with explanations of why the recommendations are unsafe could be utilized to test the system’s understanding of why specific texts are dangerous.

Acquiring textual examples of unsafe scenarios on the internet is challenging due to the intricacies involved in identification. No explicit keywords or known language patterns can be used to automate the process of finding covertly unsafe text. However, several websites with communities focused on offering advice, such as Reddit or Twitter, may be a good starting place for locating recommendations that lead to potentially unsafe outcomes. The data annotation process may also prove challenging as covertly unsafe text spans several different knowledge domains. As a result, a collaboration between crowd workers and domain experts would likely be most effective for the annotation process. Domain experts can provide in-depth knowledge, while crowd workers can provide diverse perspectives, and when combined, this provides the most coverage for various covertly unsafe scenarios.

4.2 Creating Safe Systems

To mitigate covertly unsafe text within systems, we focus on three threads: system knowledge (§4.2.1), controlled text generation (§4.2.2), and explainability (§4.2.3). These threads directly connect (Figure 3) to our categorizations of covertly unsafe text (§3) and provide promising directions toward mitigating covertly unsafe text. Note that this set of topics is not comprehensive, and we encourage researchers to explore further directions.

4.2.1 Integrating System Knowledge

A system’s access to relevant knowledge, whether commonsense or domain-specific, is critical for text safety. The system requires external knowledge to recognize the physical harm caused for language within the limited information category. Understanding the connections and contradictions between various actions can help to prevent generating text in the incompatible information category. Additionally, access to factual knowledge can avoid generating incorrect information.

One solution to make commonsense-aware systems is to use a knowledge base. This benefit is that information on an extensive range of topics can be consolidated and used to augment NLG models. Several studies have focused on creating knowledge bases that encode general human knowledge about the world (Speer et al., 2017; Sap et al., 2019; Zhang et al., 2020). Although they contain valuable knowledge for many systems, they do not emphasize common concepts related to human safety. As such, there is potential to better target the problem of covertly unsafe text through a commonsense knowledge base specifically focused on human safety knowledge. For example, leveraging a knowledge graph with actions and physical effects by adding safe and unsafe relations can help make safety more explicit. If these graphs can also model interactions between multiple actions, they can help prevent incompatible information.

Systems requiring specific knowledge related to certain topics can benefit from domain-specific knowledge. For example, a medical chatbot can utilize a medical knowledge base to ensure that there are no gaps in specialized knowledge (Bodenreider, 2004), as well as account for user-specific circumstances. Medical applications may also utilize systems that model the interactions between various actions and medications (Hester et al., 2011). Conversational agents that are targeted to specific

domains can use a pre-determined domain-specific vocabulary (Choudhary et al., 2017) or domain-specific knowledge triples (Zhu et al., 2017). Systems with domain contextualized information that also integrate safe and unsafe relations can be particularly effective in mitigate covertly unsafe text. A factual knowledge base can also help prevent generating false information or fact-check generated claims (Thorne et al., 2018; Jiang et al., 2020).

In addition to knowledge bases, several benchmarks exist for tasks related to commonsense reasoning (e.g., Gordon et al. (2012); Mostafazadeh et al. (2016); Zellers et al. (2018)) to gauge a system’s general commonsense reasoning abilities. However, they may not accurately depict a model’s reasoning ability in safety-critical scenarios. As a result, there is a need for formulating more safety-specific commonsense reasoning tasks. Consider the proposed safety datasets (§4.1); one possible task could be to determine the physical effect of an unsafe statement, which would test a system’s causal reasoning capabilities.

4.2.2 Controlled Text Generation

A fundamental aspect of natural language generation is controllability, the ability to enforce constraints on generated text. Controlled Text Generation (CTG) can naturally apply to text safety by preventing the generation of covertly unsafe text. Previous research on controllable text generation methods for large pre-trained language models has focused on controlling sentiment, topic, persona, or keywords (Zhang et al., 2022). However, establishing constraints for unsafe text and adapting this to existing CTG methods is not trivial because covertly unsafe text spans many domains.

Fine-tuning is one method of producing controlled text (Devlin et al., 2019), which has already been applied to toxicity (Solaiman and Dennison, 2021) and can be an approach adaptable to other safety-related systems. For instance, a question-answering system can be fine-tuned on a dataset for text safety (§4.1) to adapt the system to such text. Furthermore, reinforcement learning approaches to fine-tuning help incorporate human judgments and preferences into development (Ziegler et al., 2019; Bai et al., 2022), which can help mitigate biases.

Prompting prepends additional context to the input of a task for a model to condition on during generation (Askell et al., 2021). These prepended trigger words can help prevent systems from generating incorrect information. For instance, masked

language models can control text generation to only factual knowledge (Shin et al., 2020) or toxic and unsafe responses adversarially (Wallace et al., 2019). Applying this to safety, we can prompt systems with statements like “respond to the query with a safe response.” Similarly, prefix-tuning can also replace fine-tuning (Li and Liang, 2021).

Another less computationally intensive option is post-processing, which does not involve modifying model parameters. One simple approach uses attribute classifiers combined with large pre-trained language models, allowing text to be generated conditioned on various attributes like topic or sentiment (Dathathri et al., 2019); attribute classifiers can be applied to safe text generation for safe and unsafe text classes. Other decoding algorithms use predicate logic constraints or lookahead heuristics, which may be useful for preventing unsafe text from occurring in the generated output (Lu et al., 2020, 2021). Additionally, lexically constrained decoding can be utilized to promote the generation of factual information (Mao et al., 2020).

Faithfulness. This subset of CTG focuses on preventing hallucinating new information, measured by how accurately an explanation of a model reflects its actual reasoning (Jacovi and Goldberg, 2020). Thus, a system would be considered unfaithful if the explanation does not match the decision or if similar inputs and outputs receive vastly different explanations (Jacovi and Goldberg, 2020). Predictive uncertainty between similar inputs and generated outputs can also correspond with occurrences of hallucinations (Xiao and Wang, 2021).

Faithfulness, as a result, can directly correlate to incorrect covertly unsafe text (§3.3) because deviating from accurate information can incorporate error and produce results that may lead to physical harm. For example, a throat-soothing remedy recommendation to drink 100°F hallucinated to 100°C water can turn soothing warm water into scalding hot burns. One method to develop faithful and safe systems can be to evaluate generated text by comparing it with a system’s safety-oriented knowledge base (§4.2.1) to prevent hallucinations and ensure text safety.

4.2.3 Explainability

Explainability is the ability to justify a system’s decision based on given inputs and comes in several forms (Adadi and Berrada, 2018; Gerke et al., 2020; Davahli et al., 2021). Two flavors particularly relevant in the context of covertly unsafe text in-

clude diagnosing input-output mappings (Koh and Liang, 2017; Verma et al., 2020) and generating human-readable reasoning (Kojima et al., 2022).

Particularly in safety-critical systems, it is important to have interpretable models to understand the reasoning behind recommendations that directly impact users (Goodman and Flaxman, 2017); incorrect recommendations in these sensitive areas can lead to covertly unsafe text. For example, recommending chemotherapy on an incorrect cancer diagnosis would be considered physical harm as the patient would be exposed to cell-killing chemicals (Zhang et al., 2019).

Two common approaches to provide insights into black-box models are perturbation functions (Koh and Liang, 2017), which seek to see output differences when local inputs are tweaked, and counterfactual reasoning (Verma et al., 2020), which considers the global alternative to determine input is needed to reach such state. Counterfactuals provide the advantage of understanding the global impacts of inputs but are challenging to implement in practice; conversely, perturbation functions are more efficient but only offer insights into how local changes influence the system output.

Interpretability. Human-interpretable explanations provide reasoning to justify a system’s decisions. This is a useful way to understand black boxes and a valuable resource to diagnose systems generating covertly unsafe text. However, these generated explanations may be unsafe. For example, we can adapt a QA approach (Kojima et al., 2022) that asks for an explanation of the model’s reasoning with the question “Should I get the Shingles vaccine?” A covertly unsafe explanation would be “yes because it helps build immunity to a painful disease” since the vaccine is only safe for adults. We recommend the other mitigation strategies discussed to handle this problem.

4.3 Metrics Capturing Text Safety

The final step in the ML pipeline is to evaluate the quality of outputs in terms of safety. Using existing resources, one method is to compare the generated output to a set of safe versus unsafe text, compute the difference, and test for significance; when applied to generation and summarization tasks, common n-gram metrics such as ROUGE and BLEU (Lin, 2004; Papineni et al., 2002) test for exact match and may miss the sentiment. An initial approach for richer sentiments includes BERTScore

(Zhang et al., 2019), which tests for vector similarity instead. Likelihood methods like perplexity can face issues with over-reliance on the training data, which can propagate biases.

Metrics related to faithfulness evaluate factual consistency in NLG systems (Maynez et al., 2020; Alvarez-Melis and Jaakkola, 2018; Wolf et al., 2019). These metrics can help capture limited, incompatible, or incorrect information present in covertly unsafe text due to hallucinations (Li et al., 2022). Some of the best-performing methods for achieving this are entailment-based metrics involving Natural Language Inference or QA-based metrics (Honovich et al., 2022).

Beyond general evaluation metrics, there lacks an excellent safety-specific metric to capture whether texts are covertly unsafe. Fundamentally desirable qualities in any well-formed metric include optimizability by being differentiable and not compromising task performance. In the context of safety, this metric should parallel human safety judgments and, when optimized, should minimize unsafe text. One metric could capture the probability that a particular action is unsafe; another metric can align with the severity of physical harm caused, ranging from minor pains to cruel torture or death. With these safety metrics, it is also important to consider the diversity in perspectives, as different individuals and cultures may uniquely rank what is more dangerous.

4.4 Detection of Human-Written Unsafe Text

In addition to mitigating the generation of unsafe text, several of these strategies are general enough to enable the detection of AI or human-written unsafe text. For example, using explainable system approaches to an unsafe text detector can provide valuable insights as to why a specific text with incorrect information is physically unsafe. Similarly, datasets for text safety can be adapted for detection settings by building a safety classifier instead. Detection systems are directly applicable to communities of discourse where unsafe text may circulate. Yet, our work does not focus on detecting unsafe text due to potential censorship issues and encourages future researchers to explore this delicate balance.

5 An Interdisciplinary Path to Safe AI

So far, our discussion has been focused on technical solutions to prevent AI systems from generating covertly unsafe text. As harm is a sensitive



Figure 4: Interdisciplinary steps toward mitigating physical harm caused by covertly unsafe text.

topic with many legal repercussions, we will now ground our discussion of physical harm on how current policy interacts with harmful AI. We also consider human factors that are out of scope for current AI systems, including foreseeability, target, and motive; we evaluate how these may apply in the detection context and call for an interdisciplinary approach to tackle these issues (Figure 4). This approach can effectively mitigate bias against protected groups that may be susceptible targets.

5.1 Interactions of Harmful AI and Policy

Policy frameworks for addressing harmful AI are in early development. In its absence, principles for AI safety are likely to be developed piecemeal by courts that hold stakeholders associated with AI systems liable for harm under existing tort⁸ laws.

Applying existing liability principles to intelligent systems presents complex challenges. Legal scholars disagree about the applicability of the extant liability regime (Padovan et al., 2022) since standard concepts in liability do not apply to AI straightforwardly (Villasenor, 2019).

One compelling problem is assessing producers’ duty to foresee harm their AI systems produce. Foreseeability is central to how courts assign responsibility for harm; when such a case arises, courts will consider whether the system producers could have anticipated the harm and taken steps to prevent it (Selbst, 2020; Giuffrida, 2019). For personal assistants, foreseeability declines with increased degrees of separation concerning physical harm (§2). However, despite covertly unsafe text being less foreseeable than overtly unsafe text, it still poses a danger to users of intelligent systems, and this problem needs to be equally prioritized by system producers. Because of these dangers, policymakers should also dive deeper into these issues to develop standards for addressing different degrees of physically harmful text.

⁸relating to negligence

5.2 Human Involvement in the ML Pipeline

Integrating a human-centered approach is necessary to address covertly unsafe text most effectively. A purely automated solution can miss the social context needed to address the human-centered issue of safety (Ehsan et al., 2021). Factors such as target and motive can raise other regulatory concerns if intelligent systems foster malicious behavior; a profiling system that outputs covertly unsafe text to trick children into consuming dangerous substances would be a prime example.

Task Creation. When creating new tasks, they tend to be constructed to match humans’ definition of success. This is generally positive in the context of safety as humans tend to have a strong understanding of danger; yet, this can be negative as humans tend to take knowledge for granted, not assumed by a model. This gap in system knowledge may create unsafe models when a susceptible group also does not have that tacit knowledge that individuals with more domain expertise in that particular area. For example, suppose someone encounters an unknown powder. An instinct and recommendation may be to identify it using the five senses. Still, those with more domain expertise may assume it is dangerous and contact the authority instead. To mitigate potential disparities, we encourage constructing focus groups for a variety of backgrounds to review new safety-related tasks and metrics. This would minimize incorrect assumptions and maximize coverage of the different types of covertly unsafe physical harm.

Crowd Sourcing. Crowd workers are likely involved in many stages of the pipeline, from helping to write context to unsafe scenarios to human evaluation of the safety of generated texts. Like task creation, crowd workers may have unique perceptions of safety influenced by their backgrounds and beliefs (Sap et al., 2021). As a result, it is ideal to go beyond a simple convenience sample and acquire crowd workers with diverse perspectives to help mitigate biases that may span from perceptions of safety. For future research, this can be expanded to explore different definitions of safety.

5.3 Bridging Gaps with Social Workers

Social workers can bridge the gap between impacted communities, computer scientists, and policymakers. Since social workers are often immersed in marginalized communities (Mathiyazhagan et al., 2021), they can help computer scientists

and policymakers understand different user groups and impacted communities, providing critical feedback on defining, measuring, and mitigating unsafe language from human-written or machine-generated text. Furthermore, social workers can help educate these communities to exercise caution when interacting with intelligent systems or machine learning models, as system outputs may not necessarily be truthful or safe. Social workers understand the cultural backgrounds of minority communities and can provide insight into misunderstandings or situations in which misinformation may be more likely to be accepted. A collaboration between domain experts and social workers can further benefit communities by advising on the risks of unsafe situations.

6 Conclusion

In this paper, we address increasing concerns over text safety. We first establish degrees of separation with respect to physical harm as a methodology to label physically unsafe text as either overtly, covertly, or indirectly unsafe. We further dissect covertly unsafe text with the cause of either limited, incompatible, or incorrect information. Each type of covertly unsafe text has unique attributes requiring different strategies to resolve; we discuss these methods with respect to the ML pipeline to provide future researchers inspiration to tackle the issues of text safety. Finally, we discuss an interdisciplinary approach to mitigating covertly unsafe text.

Covertly unsafe text is a challenging problem that spans a breadth of domains with no overtly unifying common thread. Since covertly unsafe text is subtle yet equally dangerous to overtly unsafe text, we argue that stakeholders and policymakers must acknowledge and proactively prioritize it to protect users' physical safety when interacting with intelligent systems.

Limitations

While our research touches upon physical harm, our paper primarily discusses covertly unsafe text, limiting the discussion of other types of physically harmful text, including overtly unsafe and indirectly unsafe text. While the latter types of unsafe text are equally problematic in causing physical harm, our paper does not focus on either of these aspects due to the expansive coverage of previously existing research on these topics.

In addition to limitations in the spectrum of physically harmful text, our work may be limited in categorizing covertly unsafe text. We provide sub-categories of limited, incompatible, and incorrect information that causes text to be covertly unsafe, but these categories may not be comprehensive.

This research aims to address the problem of covertly unsafe text and inspire future researchers to help improve intelligent systems by exploring ways to tackle this challenging problem. We encourage readers to consider the problem space of covertly unsafe text, whether there may be additional categorizations of these texts, and even propose new mitigation strategies.

Ethical Considerations

We acknowledge that our research touches upon sensitive topics of harm that affect individuals differently. Our work discusses commonsense and categorizations of harm with a singular definition of safety in an attempt to improve text safety universally, yet we note that personal backgrounds influence and shape people's views and values non-uniformly, which can affect people's perceptions of harm and safety differently. As a result, bias may propagate through efforts to improve text safety, which can impact protected groups disproportionately. We encourage researchers in this area to be aware of these potential factors and proactively attempt to mitigate bias against protected groups by applying a conscious human-centered strategy.

Acknowledgements

We thank our reviewers for their helpful feedback. We also thank Rukmini Bapat for her early contributions in the initial literature search. We would also like to thank Amazon AWS Machine Learning Research Award and Amazon Alexa Knowledge for their generous support. This material is based upon work supported in part by the National Science Foundation under Grant #2048122. The authors are solely responsible for the contents of the paper, and the opinions expressed in this publication do not reflect the official policy or position of the funding agencies. We also thank the Robert N. Noyce Trust for their generous gift to the University of California via the Noyce Initiative.

References

- Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.
- Abdulaziz Alamri and Mark Stevenson. 2015. Automatic identification of potentially contradictory claims to support systematic reviews. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 930–937.
- David Alvarez-Melis and Tommi S. Jaakkola. 2018. [On the robustness of interpretability methods](#).
- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. [A general language assistant as a laboratory for alignment](#).
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson 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 Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#).
- John A. Bateman. 1990. Upper modeling: A general organization of knowledge for natural language processing.
- 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*, pages 610–623.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.
- Luke Breitfeller, Emily Ahn, Aldrian Obaja Muis, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *EMNLP*.
- Melissa Carmona. 2022. [Melatonin and xanax — The Recovery Village](#). [Online; accessed 16-June-2022].
- Erin Carson. 2021. [Milk crate challenge: Why people are taking huge, terrifying falls on social media — CNET](#). [Online; accessed 16-June-2022].
- Carlos M Castorena, Itzel M Abundez, Roberto Alejo, Everardo E Granda-Gutiérrez, Eréndira Rendón, and Octavio Villegas. 2021. Deep neural network for gender-based violence detection on twitter messages. *Mathematics*, 9(8):807.
- CBS News. 2013. [Cinnamon challenge dangerous to lungs, new report warns — CBS News](#). [Online; accessed 16-June-2022].
- Serina Chang, Ruiqi Zhong, Ethan Adams, Fei-Tzin Lee, Siddharth Varia, Desmond Patton, William Frey, Chris Kedzie, and Kathleen McKeown. 2018. Detecting gang-involved escalation on social media using context. *arXiv preprint arXiv:1809.03632*.
- Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Athena Vakali, and Nicolas Kourtellis. 2019. Detecting cyberbullying and cyberaggression in social media. *ACM Transactions on the Web (TWEB)*, 13(3):1–51.
- Sajal Choudhary, Prerna Srivastava, Lyle H. Ungar, and João Sedoc. 2017. Domain aware neural dialog system. *ArXiv*, abs/1708.00897.
- 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. *arXiv preprint arXiv:1912.02164*.
- Mohammad Reza Davahli, Waldemar Karwowski, Krzysztof Fiok, Thomas Wan, and Hamid R Parsaei. 2021. Controlling safety of artificial intelligence-based systems in healthcare. *Symmetry*, 13(1):102.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.
- Emily Dinan, Gavin Abercrombie, Ari Bergman, Shannon L. Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. Safetykit: First aid for measuring safety in open-domain conversational systems. In *ACL*.
- Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. [Expanding explainability: Towards social transparency in ai systems](#). In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA. Association for Computing Machinery.
- Lynne Eldridge. 2021. [The link between nicotine and cancer — Verywell Health](#). [Online; accessed 16-June-2022].
- Sara Gerke, Timo Minssen, and Glenn Cohen. 2020. Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare*, pages 295–336. Elsevier.

- Iria Giuffrida. 2019. Liability for ai decision-making: some legal and ethical considerations. *Fordham L. Rev.*, 88:439.
- Gregorio Arturo Reyes González and Francisco J Cantu-Ortiz. 2021. A sentiment analysis and unsupervised learning approach to digital violence against women: Monterrey case. In *2021 4th International Conference on Information and Computer Technologies (ICICT)*, pages 18–26. IEEE.
- Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.*, 38:50–57.
- Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of common-sense causal reasoning. In **SEMEVAL*.
- Anne Helmenstine. 2020. [Bleach and alcohol make chloroform – why you shouldn't mix disinfectants — Science Notes](#). [Online; accessed 16-June-2022].
- Robert L. Hester, Alison J. Brown, Leland D. Husband, Radu Iliescu, Drew Pruett, Richard L. Summers, and Thomas G. Coleman. 2011. Hummod: A modeling environment for the simulation of integrative human physiology. *Frontiers in Physiology*, 2.
- Or Honovich, Roei Aharoni, Jonathan Herzog, Haggai Taitelbaum, Doron Kukliansky, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Y. Matias. 2022. True: Re-evaluating factual consistency evaluation. In *DIALDOC*.
- Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Kumar Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. In *FINDINGS*.
- David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A just and comprehensive strategy for using nlp to address online abuse. *arXiv preprint arXiv:1906.01738*.
- Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *ArXiv*, abs/2012.12305.
- Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. *ArXiv*, abs/1703.04730.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.
- Sharon Levy, Kevin Mo, Wenhan Xiong, and William Yang Wang. 2021a. [Open-Domain question-Answering for COVID-19 and other emergent domains](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 259–266, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sharon Levy, Michael Saxon, and William Yang Wang. 2021b. [Investigating memorization of conspiracy theories in text generation](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4718–4729, Online. Association for Computational Linguistics.
- Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *ArXiv*, abs/2203.05227.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *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)*, abs/2101.00190.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. 2021. Neurologic a*esque decoding: Constrained text generation with lookahead heuristics. *ArXiv*, abs/2112.08726.
- Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. [Neurologic decoding: \(un\)supervised neural text generation with predicate logic constraints](#). *CoRR*, abs/2010.12884.
- Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. 2020. Constrained abstractive summarization: Preserving factual consistency with constrained generation. *ArXiv*, abs/2010.12723.
- Siva Mathiyazhagan, Shana Kleiner, and Desmond U. Patton. 2021. [Social work in data science: Tech policy gaps and addressing harm](#).
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. *ArXiv*, abs/2005.00661.
- N. Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*.

- Javier Osorio and Alejandro Beltran. 2020. Enhancing the detection of criminal organizations in mexico using ml and nlp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.
- Paulo Henrique Padovan, Clarice Marinho Martins, and Chris Reed. 2022. Black is the new orange: how to determine ai liability. *Artificial Intelligence and Law*, pages 1–35.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Desmond Upton Patton, Kathleen McKeown, Owen Rambow, and Jamie Macbeth. 2016. Using natural language processing and qualitative analysis to intervene in gang violence: A collaboration between social work researchers and data scientists. *arXiv preprint arXiv:1609.08779*.
- Ellie Pavlick, Heng Ji, Xiaoman Pan, and Chris Callison-Burch. 2016. The gun violence database: A new task and data set for nlp. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1024.
- Sarah Masud Preum, Md. Abu Sayeed Mondol, Meiyi Ma, Hongning Wang, and John A. Stankovic. 2017. Preclude: Conflict detection in textual health advice. *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 286–296.
- Lee Rainie, Janna Quitney Anderson, and Jonathan Albright. 2017. The future of free speech, trolls, anonymity and fake news online.
- Ehud Reiter, Rohan K. Robertson, and Somaya-julu Gowri Sripada. 2003. Acquiring correct knowledge for natural language generation. *J. Artif. Intell. Res.*, 18:491–516.
- Semiu Salawu, Yulan He, and Joan A. Lumsden. 2020. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 11:3–24.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *ArXiv*, abs/1811.00146.
- 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.
- Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.
- Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Andrew D Selbst. 2020. Negligence and ai’s human users. *BUL Rev.*, 100:1315.
- Patty Osborne Shafer. 2022. General first aid for seizures — Epilepsy Foundation. [Online; accessed 16-June-2022].
- Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Eliciting knowledge from language models using automatically generated prompts. *ArXiv*, abs/2010.15980.
- Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *ArXiv*, abs/1612.03975.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *NAACL*.
- Sahil Verma, John P. Dickerson, and Keegan E. Hines. 2020. Counterfactual explanations for machine learning: A review. *ArXiv*, abs/2010.10596.
- John Villasenor. 2019. Products liability law as a way to address ai harms. *Brookings Report*.
- Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.
- Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *ALW@ACL*.
- Lior Wolf, Tomer Galanti, and Tamir Hazan. 2019. A formal approach to explainability. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’19, page 255–261, New York, NY, USA. Association for Computing Machinery.
- Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. *ArXiv*, abs/2103.15025.

Yubo Xie and Pearl Pu. 2021. How commonsense knowledge helps with natural language tasks: A survey of recent resources and methodologies. *ArXiv*, abs/2108.04674.

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.

W. Yu, Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiteng Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Computing Surveys (CSUR)*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing ki Leung. 2020. Aser: A large-scale eventuality knowledge graph. *Proceedings of The Web Conference 2020*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert](#).

Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible end-to-end dialogue system for knowledge grounded conversation. *ArXiv*, abs/1709.04264.

Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593.