NLP Security and Ethics, in the Wild

Heather Lent[1] **Erick Galinkin**[2] **Yiyi Chen**[1]
Jens Myrup Pedersen[1] **Leon Derczynski**[2,3] **Johannes Bjerva**[1]
[1]Aalborg University, Denmark
[2]NVIDIA Corporation, USA [3]IT University of Copenhagen, Denmark
{hcle, jbjerva}@cs.aau.dk

Abstract

As NLP models are used by a growing number of end-users, an area of increasing importance is NLP Security (NLPSec): assessing the vulnerability of models to malicious attacks and developing comprehensive countermeasures against them. While work at the intersection of NLP and cybersecurity has the potential to create safer NLP for all, accidental oversights can result in tangible harm (e.g., breaches of privacy or proliferation of malicious models). In this emerging field, however, the research ethics of NLP have not yet faced many of the long-standing conundrums pertinent to cybersecurity, until now. We thus examine contemporary works across NLPSec, and explore their engagement with cybersecurity's ethical norms. We identify trends across the literature, ultimately finding alarming gaps on topics like harm minimization and responsible disclosure. To alleviate these concerns, we provide concrete recommendations to help NLP researchers navigate this space more ethically, bridging the gap between traditional cybersecurity and NLP ethics, which we frame as ''white hat NLP''. The goal of this work is to help cultivate an intentional culture of ethical research for those working in NLP Security.

1 Introduction

Securing large language models (LLMs) and NLP technology in general has not been a priority until recently. Yet mass adoption of this technology has led to deployment in contexts where security failures present a risk to individuals, organizations, and society at large—demonstrated by, *inter alia*, LLM-assisted identity fraud (Ackerman, 2022), phishing campaigns (Hazell, 2023), and automated influence operations (Goldstein et al., 2023). If the latest NLP technologies are to withstand an increasing barrage of threats, NLP practitioners must now educate themselves on cybersecurity and cybersecurity practitioners must educate themselves on NLP. Yet in this process, as one culture of research adapts to another, there is potential for long-standing intra-community norms to be lost in translation, including ethical norms. In interdisciplinary research, an accidental oversight of ethical norms from one field can risk *reintroducing* previously resolved ethical dilemmas.

Discussions around ethical research conduct have long been a concern for both cybersecurity (Molander and Siang, 1998; Himma and Tavani, 2008; Matwyshyn et al., 2010; Bailey et al., 2012; Christen et al., 2020; Kohno et al., 2023) and for NLP (Wiener, 1960; Samuel, 1960; Dennett, 1997; Moor, 2006; Anderson and Anderson, 2007; Hovy and Spruit, 2016; Leidner and Plachouras, 2017), but given that these disciplines have historically been disjoint fields, the specific ethical and sociocultural norms of both fields have developed in separate silos. To better understand how interdisciplinary NLPSec has adapted to the ethical norms and values across both disciplines, we examine a set of peer-reviewed NLPSec publications from NLP venues to gauge the compliance with norms in cybersecurity. We find that several principles regarded by the cybersecurity community as best practices have not been widely adopted in NLPSec research, despite measures in the NLP peer-review process to improve research practices. Simultaneously, we find that NLP ethical norms regarding lower-resourced languages are at risk of being overlooked in NLPSec. To help save NLPSec the potential growing pains of reinventing the wheel, in this work, we seek to address this issue through an interdisciplinary conversation about ethics, with the goal of cultivating a culture of *white hat NLP*.

Ethical NLP Security (NLPSec) In cybersecurity, a ''white hat'' hacker is typically a professional hired by a company with the specific

Transactions of the Association for Computational Linguistics, vol. 13, pp. 709–743, 2025. https://doi.org/10.1162/tacl_a_00762
Action Editor: Trevor Cohn. Submission batch: 6/2024; Revision batch: 12/2024; Published 7/2025.

purpose of maintaining or increasing the existing security of a computer system. White hat hackers,[1] referred to more generally as ''ethical hackers'', are keen to engage with vendors—either their employer or a third party—whose products they have found flaws in. They aim to get security issues fixed quickly and release details of their findings to the public so defenders can evaluate their own environments and prioritize concomitant patching and mitigation efforts. In contrast, a ''black hat'' hacker represents a cybercriminal, and somewhere in between is the ''gray hat'' hacker, who infiltrates others' computer systems without permission, with the intention of enhancing security (Falk, 2004). This is generally mapped to similar activity in the context of LLMs, where LLM red teaming is defined as a limit-seeking, manual, and non-malicious activity (Inie et al., 2025). Outside the strict boundaries of ethical hackers, gray hat hackers rely on their own moral compass, which can lead to potentially precarious situations, like introducing the risk of authorized system intrusions (Christen et al., 2020). Alarmingly, much of the contemporary work in NLPSec is arguably similar to the above gray hat scenario (Figure 1). Due to the distributed and decentralized nature of the research landscape, most NLPSec researchers will necessarily lack a mandate from organizations to investigate security vulnerabilities inherent to models. Researchers also lack direct accountability to those most affected by the security breaches they study. While the ACM Code of Ethics[2] provides overarching guidance, it naturally cannot provide exact guidance for every ethical conundrum, leaving individual researchers to rely on their own moral compass, like a gray hat hacker. It is in this light that we aim to extend the framework of ''white hat'' to NLP. We thus define the scope of white hat NLPSec to consist of *works which are intentionally and carefully grounded in the established ethical best practices of both cybersecurity and NLP*.

Contributions We examine the current culture of research ethics in NLPSec through a survey of 80 peer-reviewed works across NLP venues, measuring their compliance with typical ethical practices from cybersecurity. To start, we introduce these ethical practices, and touch upon their relevance to works in NLPSec (Section 2). We then describe our selection process for papers included in the survey, and our annotation process identifying compliance with these cybersecurity ethical best practices (Section 3). Concretely, we find that ethical norms like harm minimization and responsible disclosure are not widely adopted in NLPSec (Section 4). We identify several challenges of adopting ethical norms from both cybersecurity *and* NLP to the field of NLPSec, and we provide concrete advice to help researchers engage in research *more ethically* than previously (Section 5). While the recommendations presented in this work cannot provide solutions to all potential ethical conundrums latent to NLPSec, the discussions in this work aim to highlight the urgency for discussion on research ethics in NLPSec and serve as a conversation starter.

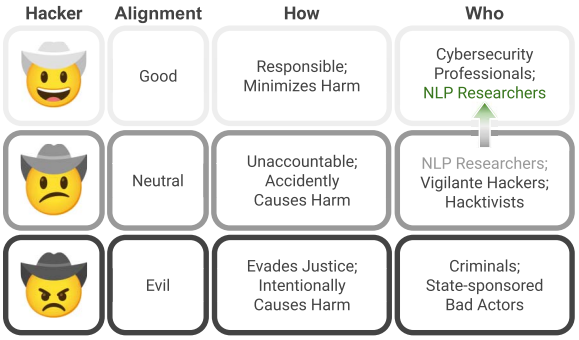

Figure 1: We argue current works across NLPSec occupy a gray area in research ethics, comparable to the cybersecurity concept of a ''gray hat'' hacker. In this work, we provide concrete recommendations to help move the field of NLPSec towards more ethically grounded research practices.

2 The Culture of Ethics in Cybersecurity

Within any field, some research may have negative consequences, and foreseeable risks generally serve as a guiding force for calibrating ethical norms. This holds particularly true for cybersecurity, as a field that is largely engaged with preempting and outmaneuvering criminals while protecting the public. As the field is naturally situated in a sensitive context, cybersecurity's research culture has unsurprisingly evolved to be one that often faces difficult ethical discussions (Bailey et al., 2012). In this section, we will introduce three ethical norms commonly expected

[1]This verbiage has long been common parlance in cybersecurity, though it has recently been criticized for its connoted colorism. As there is no widely adopted alternative, we trust readers to accept our good faith usage of the term.

[2]https://www.acm.org/code-of-ethics.

of works across cybersecurity. For each norm, we further describe potential nuances, borrowing relevant trolley problems directly from Kohno et al. (2023). Our aim is to familiarize readers with these concepts, demonstrate how best practices in cybersecurity continue to be developed, and briefly touch upon how these concepts are relevant to works in NLPSec.

2.1 Harm Minimization

A familiar concept to cybersecurity practitioners is that the ''maliciousness'' of tools is context-dependent. It is well documented that black hat hackers often misuse legitimate software (Martin, 2017; Bailey et al., 2012) for their own purposes. For instance, remote desktop management software is commonly used to facilitate remote troubleshooting for users. However, these very same tools can be used by malicious actors to maintain a foothold in a victim network. A critical part of cybersecurity methodology is establishing normal, secure usage patterns for these tools to minimize the potential risk associated with them.

To minimize harm, one must first identify the potential dangers. For example, Singla et al. (2023) aim to examine the impact of the ongoing Russia-Ukraine war on Ukrainian critical infrastructure by scanning web traffic. They observe that one potential risk inherent to such work includes the accidental disruption of end systems, which could exacerbate the current plight of non-combatant Ukrainians under the war. To minimize harm, Singla et al.'s (2023) methodology ensures uninterrupted web traffic, allows end-users to opt-out of scans, and safeguards sensitive data recovered by the scan, meanwhile collaborating with an NGO to help with responsible disclosure to Ukrainian authorities before publication. Here, potential victims of cyberattacks are protected to the greatest extent possible, largely through strategic choices laid out in the methodology.

While causing some amount of harm—however inadvertent—may be inevitable, grappling with these potential ethical considerations early on can help researchers to avoid ethical dilemmas and minimize that harm. To this end, Kohno et al. (2023) propose a security trolley problem: in the context of an AI-based employment software tool, if a data breach compromised sensitive user data, should security researchers study the leaked data (prioritizing a potential benefit to the public, if unfair bias can be established from the data) or not study it (prioritizing the affected users' right to privacy)? Kohno et al. (2023) demonstrate that both conclusions can be justified through moral philosophy frameworks, specifically consequentialist and deontological ethics,[3] respectively. The goal of this exercise is not to encourage moral relativism; on the contrary, it is to stress the necessity for continuous and principled conversations on research ethics in cybersecurity. This discussion reinforces the notion that much research in cybersecurity will demand researchers to assess not only the potential harm inflicted by their work, but also the potential harm inflicted by foregoing a study. In the absence of strict governing boards, who grant permission to researchers for specific studies, it remains the duty of an ethical researcher to remain vigilant to this characteristic of cybersecurity research and adapt accordingly. Other common strategies for harm minimization include anonymizing published datasets (Mirsky et al., 2016), limiting what information is published (Burstein, 2008), and foregoing a study entirely (Macnish and Van der Ham, 2020). Ultimately, these strategies reflect a culture of research ethics that has been developed over time.

Similar to cybersecurity, the list of potential harms considered in NLPSec can range from immediate misfortunes (e.g., models inverted to reveal private data or prompt injection leading to remote code execution) to systemic harms associated with AI systems (e.g., LLMs being weaponized to generate hate speech about a marginalized group). We explore in detail the potential risks identified by NLPSec researchers in Section 4.1 and the implications of harm minimization in NLPSec in Section 5.2.

2.2 Coordinated Vulnerability Disclosure

In traditional cybersecurity, a key aspect of ''white hat'' ethical hacking is the clear and timely disclosure of relevant information through a process known as coordinated vulnerability disclosure (CVD) (ISO 29147:2018). In CVD, when an issue is found, it is reported to the vendor on a

[3]In Kohno et al. (2023), utilitarianism is adopted to conduct consequentialist analyses, centering the outcome of an action, whether it produces the greatest net positive well-being; the deontological analyses follow Kantian moral philosophy, whereby humans, as rational beings, have an absolute moral duty to justice regardless of consequences.

best-effort basis *before* findings are published. This gives those responsible for the affected software an opportunity to address the discovered issue ahead of any public disclosure, often on some agreed-upon timeline. When the vulnerability is disclosed by a vendor, researcher, or in a joint release, information about remediation or prevention is included so affected parties can minimize or mitigate their exposure to the issue. To this end, CVD offers security researchers a way to maintain transparency, without assisting malicious actors, as the published vulnerabilities will hopefully have already been patched.

In the real world, of course, CVD can be complicated by a variety of factors. For example, Kohno et al. (2023) detail another pertinent trolley problem: imagine an industrial researcher is assigned to review an anonymous manuscript, which reveals a severe security vulnerability in software supported by the industrial researcher's employer. Should the researcher disclose the security vulnerability to their employer (thus prioritizing the safety of a large number of end-users, who would otherwise be exposed) or not disclose it (thus prioritizing the authors' rights with respect to peer-review)? As pointed out by the Menlo Report (Bailey et al., 2012), in such circumstances, different stakeholders are likely to have different priorities, and sometimes the most ethical action may be in direct conflict with one's best interests. Accordingly, the above trolley problem can be reexamined from the perspective of different stakeholders to even further explore the ethical consequences of one decision over another. Perhaps this ethical conundrum could have been avoided entirely, if the imagined authors had planned to do CVD from the start. In most cases, however, cybersecurity practitioners are highly incentivized to complete CVD. From bug bounty programs to vulnerability disclosure leaderboards, CVD is a well-established norm, which helps companies and organizations secure these systems, while being highly prestigious for researchers.

In the context of NLPSec, at face value, CVD is relevant to works examining security vulnerabilities of proprietary language technologies. We discuss the presence and ramifications of CVD in NLPSec in Sections 4.2 and 5.3, respectively.

2.3 Public Disclosure

Another common practice in cybersecurity as part of, or (less desirably) in lieu of, CVD is public disclosure. The essential idea is that attackers, particularly those motivated by criminal or national security desires, are incentivized to take a potentially interesting or useful vulnerability and uncover ways to exploit it. In contrast, defenders are generally already busy triaging alerts, responding to incidents, managing defensive infrastructure, and other important tasks that constrain their ability to develop a way to test for and detect potentially vulnerable systems. In other words, defenders, stymied by other responsibilities, may move at a much a slower pace than attackers. As reported by Rapid7 (2022), more than half of widely exploited vulnerabilities were leveraged by attackers against victims in less than one day after disclosure. This puts defenders at a distinct disadvantage in terms of the speed at which they can react to emerging threats, placing the fate of their security in the hands of software vendors and leaving them at the mercy of their own scheduled patching cycles. In this way, defenders can be informed of potential threats quickly and benefit greatly from the help of white hat hackers who share their findings in a responsible manner.

One such example where public disclosure may be called for, in lieu of CVD, is in the case a vulnerability is identified in relation to a company that no longer exits and so there is no entity with which to coordinate. Such a scenario is not unthinkable in the real world, as Kohno et al. (2023) introduce another security trolley problem whereby security researchers have identified a vulnerability in an imagined medical device that is embedded in sizable population of patients but is no longer serviceable, as the manufacturer is out of business. Should the researchers publicly disclose the vulnerability (thus prioritizing the patients' right to informed consent and bodily autonomy) or not disclose it (prioritizing the patients' peace-of-mind and happiness)? Again, Kohno et al. (2023) demonstrate that both decisions can be reached via different frameworks of moral philosophy; the purpose of this illustration is equivocally *not* to show that any decision taken by the researchers can be simply justified after the fact, but rather to again underline the importance of a cybersecurity research landscape, which is actively engaged in conversations on ethics. In practice, similar scenarios to the above trolley problem have led to the formation of Computer Emergency Response Teams (CERTs). If a practitioner identifies a bug which has no obvious

path for CVD, and the bug is likely to be abused by malicious actors upon public disclosure, they may opt to disclose the bug only to a CERT or to a trusted community. The creation of such CERTS or trusted communities in critical spaces again underscores that the act of disclosure is highly important and carefully considered across cybersecurity.

In NLPSec, publication in open-access venues such as the ACL or ArXiv is a form of public disclosure. Like defenders in cybersecurity, the majority of NLPSec researchers have limited resources, putting them at a relative disadvantage to bad actors. In this way, researchers can help each other by open-sourcing code, where appropriate. We examine the prevalence of fully open-source works in Section 4.3 and revisit public disclosure in Section 5.1.

3 Methodology

We examine 80 pertinent, peer-reviewed works across NLPSec for common trends and themes pertaining to discussions on research ethics. All papers were manually gathered from the ACL Anthology,[4] by querying keywords associated with common attack types (i.e., ''security'' + {''adversarial'', ''backdoor'', ''data recreation'', ''inversion'', ''instance encoding''}) to ensure they fall within the scope of NLPSec. We specifically did not include ''attack'' or ''defense'' in our keyword search, to avoid influencing the results. For each keyword search, we examined the relevance-sorted list of results. As terms like ''adversarial'' or ''inversion'' can be used in a wide variety of contexts beyond NLPSec, we first review each paper title, and keep those which are obviously relevant to NLPSec; where relevance is unclear from the title alone, we further scan the abstract to determine relevance. Papers dated between 2019 and 2023 were obtained from the anthology in January 2024 (n = 60). To include papers published in 2024, we used the same procedure in November 2024, following the publication of the EMNLP 2024 proceedings (n = 20). Accordingly, all 80 papers for this study are guaranteed to be relevant and peer-reviewed. While a substantial number of publications in NLPSec—particularly those concerning attack methods—are published in preprint archives like ArXiv, such papers do not necessarily go through peer review (e.g., Zou et al., 2023; Mehrotra et al., 2023). We intentionally limit ourselves to peer-reviewed publications to ensure rigorous publications whose claims have been vetted, rather than considering preprints whose claims we must take at face value. Additionally, publications in major conferences and journals are obliged to abide by the ACM Code of Ethics; as such, we assume good intent by the authors to meet a standard of ethics that is acceptable to the scientific community. For each of the 80 papers, we manually annotate the following:

Attack Scenario. (Values: `Adversarial`, `Backdoor`, or `Data Reconstruction attack`). This is coded in accordance with the keyword which was used to retrieve the paper.

Main Contribution. (Values: `Attack`, `Defense`, or `Both`). We assign this value based on the text of the title and abstract only. For example, a paper which discusses exclusively an attack method in the title and abstract is coded as `Attack`, even if a defense is offered later in the paper, for example in the list of contributions, or in an analysis. In this way, our coding process intends to mirror the authors' framing of their own work.

Discussion of Ethical Concerns. (Values: `Yes`, `No`). Here, we first look for the presence of a dedicated ethics section; if the paper has one, it is coded `Yes`. If the paper does not have such a dedicated section, we continue to search for discussions on ethics by looking for a broader impacts section, then reading the conclusion, the limitations, and the introduction, as these sections typically contain high-level reflections on topics such as ethics. If a discussion of ethics has still not been identified in the aforementioned sections, we finally search the document for the lemma ''ethic'' and examine all possible matches to determine whether the paper discusses ethics. Otherwise, if there are no matches, the paper is coded as `No`.

Dual Use and Misuse. (Values: `Yes`, `No`). The coding process for dual use and misuse occurs in step with that of the ethics discussion. That is, we first check the ethics section (if it exists), as this is where an outright dialogue on misuse is most likely to occur. If we do not find it there, we then check the conclusion, limitations, and

[4] `https://aclanthology.org/`.

introduction, accordingly. If we still have not located discourse on dual use or misuse in these sections, we search the full document for the lemmas ‘‘use’’, ‘‘leverage’’, and ‘‘malicious’’, and check the context of any matches. If no discussion has been identified through this process, the paper is coded `No`. Note that we annotate dual use as misuse separately (definitions in Section 5.1).

Coordinated Vulnerability Disclosure. (Values: `Yes`, `No`). For this variable, we check the introduction, conclusion, ethics/broader impact, limitations, and also footnotes. If CVD is not identified, we search for the following lemmas: ‘‘disclose’’, ‘‘contact’’, ‘‘reach’’, ‘‘communicate’’, and ‘‘company’’, in an attempt to locate discussions on coordinated vulnerability disclosure. If still nothing is found, the paper is coded `No` for CVD.

Open-Source Code. (Values: `Yes`, `Empty`, `No`). Open-source papers typically link to the project's Github page in a footnote on the first page. Accordingly, we search for ‘‘github’’, and cross-reference any repositories with the associated footnote, to confirm that the linked code is contributed by the authors. When a Github repository is found, we follow this link to examine the availability and contents of the repository. If the link is broken, we code the value `Empty (Broken Link)`. If the repository contains only a README and no scripts, we code it as `Empty (Empty Repository)`. When a repository cannot be located through the footnotes, we further search the following lemmas: ‘‘code’’, ‘‘provide’’, and ‘‘publish’’, and check the surrounding context, before ultimately coding the paper `No`, for no code.

Other Metadata. This includes author **affiliations**, **datasets** used, **languages** involved, and the **models** attacked. For each paper, we document the unique set of affiliations and their associated countries. Datasets are collected from the descriptions of experiments and corroborated directly against tables presented in a given work. Accordingly, we also record the languages present for each dataset, as described by the paper in the experiments, or otherwise inferred from the scope of the paper; most papers work solely on English alone, and those with a wider scope of languages tend to very clearly document this. Victim models are identified in the same manner as datasets, through the description of the experiments and the associated tables.

Figure 2: Half of sampled works across NLPSec include no discussion of ethics (41/80), even when introducing attacks that could be misused by bad actors, demonstrating the pressing need to assess and discuss ethics in this space.

Resulting from the annotation process, in Figure 2, we observe that approximately half (51%) of the works include no discussion on ethical considerations, underscoring the critical need for a broader dialogue on research ethics in NLPSec. Note that these findings refer to our sample, covering papers vetted by reviewers which should adhere to established ethical standards. We fear that the situation beyond reputable, peer-reviewed venues may be substantially worse. For the full list of sampled papers and their associated annotations, we refer readers to Appendix A Tables 1, 2, and 3. Additionally, Appendix B provides notable metadata concerning the sampled papers, such as years and venue of publication (Figures 8 and 9), the global distribution of author affiliations (Figure 10), and the most used datasets (Figure 11).

3.1 Survey Coverage

To ensure that our sample of 80 papers is representative of works across NLPSec, we employ a citation crawler[5] through the Semantic Scholar API (Kinney et al., 2023). We begin with 11 seed papers, widely cited across NLP (i.e., Mikolov et al., 2013a,b; Devlin et al., 2019a; Sanh et al., 2019; Liu et al., 2019; Raffel et al., 2019; Workshop et al., 2023; Touvron et al., 2023a,b; Jiang et al., 2023; Achiam et al., 2023). The citation networks for the seed papers points us to 223,078 citing works (as of December 2024), which can be considered to be broadly topical to NLP. From the resulting citations, we apply additional filters, in order to further narrow the scope from NLP to NLPSec. Concretely, we check each paper's title

[5] `https://gist.github.com/hclent/`.

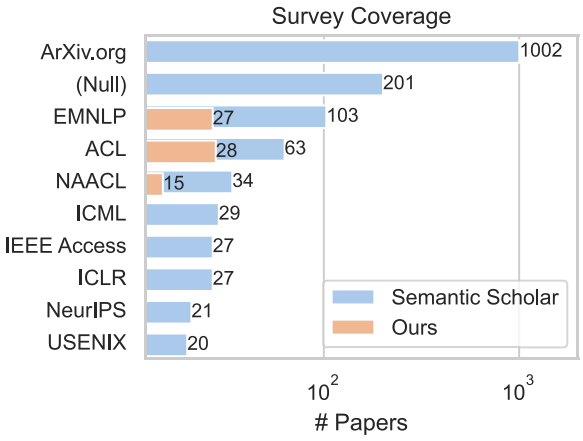

Figure 3: An approximation of the broader field of NLPSec, across the 10 most frequent venues. Where multiple venues exist for a single paper, the Semantic Scholar API prioritizes publisher venues (e.g., ACL) over preprint repositories (e.g., ArXiv), when available. A ''Null'' value is returned by the API when Semantic Scholar lacks reliable venue metadata for that paper. Against this approximation, we also show a subset of our 80 sampled papers, which specifically belong to the displayed venues (''Ours''). Here, we include Findings papers in the counts for EMNLP, ACL, and NAACL, as the API did not distinguish between them.

and abstract for matches with the following lemma: ''secur'', ''attack'', ''defen''. Again, we intentionally avoid ambiguous search terms like ''adversarial'', as they are widely used outside of an NLPSec context. Through this filtering process, we identify 2,782 unique publications, which gives us a coarse-grained estimation of the broader scope of NLPSec. These citations are dated from 2020–2024, with a reasonable share at ACL* venues. We find that 70 of our sample papers are present in the ∼200 works at ACL* venues (Figure 3), demonstrating that our survey represents upwards of 35% of works in main ACL* venues, providing a healthy sample size to draw conclusions on trends pertaining to ethics across NLPSec.

4 NLPSec Survey Results

Despite NLPSec's position as an interdisciplinary field, we find that the works in our survey have largely failed to adopt ethical norms of cybsersecurity research.

4.1 NLPSec Disagrees on Potential Harms

The attempt to minimize harm requires first an assessment of what constitutes harm, in a given

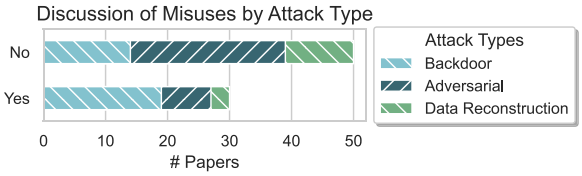

Figure 4: Across our sample, most papers do not mention or discuss the potential for misuse. No works discuss dual use.

context. We find that explorations of risk assessment vary widely across surveyed NLPSec works. Many cite *misuse*—the potential for malicious actors to weaponize their proposed methods or code towards criminal ends—as harm that could possibly arise as a result of public disclosure (e.g., Xu et al., 2021; Zeng et al., 2021; Li et al., 2023d). Though, the recognition of misuse as a potential harm varies across attack types within our sample, with works related to Backdoor attacks being the most concerned, and works related to Adversarial attacks being the least (see Figure 4). At the same time, some authors assert that there are *no* inherent risks to their work (e.g., Liu et al., 2023; Zhang et al., 2022b). Such claims of risk-free research are more often found in manuscripts introducing defense mechanisms, as the authors may mention how defenses are unlikely to be misused by the public (e.g., Qi et al., 2021a), impossible to misuse (e.g., Jin et al., 2022), or even morally noble (e.g., Li et al., 2021c). In direct contrast, Yang et al. (2021a) discuss the liability that their proposed defense mechanism could be directly studied by bad actors, wishing to sidestep such safety measures.

The above findings illuminate that **NLPSec researchers disagree on *what* potential harms may exist, and *whether* potential harms exist at all.** How can the norm of harm minimization thus be adopted in NLPSec, in the context of such nonagreement?; and who is at risk of suffering harms, if this nonagreement is left unresolved? In Section 5.1, we expand on misuse NLPSec in more detail; in Section 5.2, we discuss how current trends in NLPSec research further jeapordize vulnerable communities.

4.2 NLPSec Lacks a Culture of CVD

Among our sample of 80 works in NLPSec, we find no outright declarations of CVD. In part, this can likely be explained by the kinds of victim models researchers choose to experiment on (see

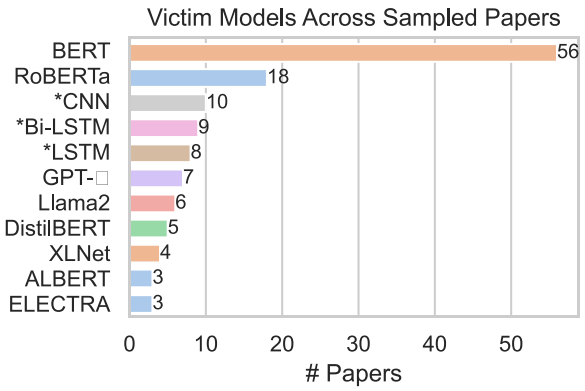

Figure 5: Distribution of victim (or suspect) models, across the sampled works in NLPSec. Models prepended with an asterisk ‘‘*’’ are ones trained individually, rather than downloaded from pre-trained weights. For the purposes of this annotation, we do not disambiguate between model sizes, for example, BERT-large versus BERT-small. Similarly the GPT-□ label represents the set of all GPT-based models across our sample (i.e., `text-embeddings-ada-002`, GPT-2, GPT-J, GPT2-XL, GPT3.5, and GPT-NEO1.3). For visualization purposes, we exclude the long tail of models that have been attacked (or defended) by only one or two papers.

Figure 5). Some are simple, self-trained models, without pre-trained weights (e.g., Bi-LSTMs), while others are long-time, staple LMs (e.g., BERT (Devlin et al., 2019b) and RoBERTa (Liu et al., 2019)), which no longer represent the state-of-the-art. Still, as models like BERT and RoBERTa are still in heavy circulation, with official versions in circulation from businesses (Google and Meta, respectively), they are candidates for CVD, at face value. Moreover, works engaging with newer, proprietary models (e.g., those from OpenAI), still do not state clearly whether or not CVD occurred as a part of the publication process. Thus it is clear that the ethical norm of CVD in cybersecurity has not yet reached the world of NLPSec, which entails alarming ramifications: **there is an opportunity for cybercriminals to weaponize the security vulnerabilities revealed by works in NLPSec**. We examine potential challenges for CVD in NLPSec in Section 5.3).

4.3 NLPSec Falls Short on Public Disclosure

While researchers may not agree on the presence or severity of hazards resulting from the publication of their work, it is generally accepted

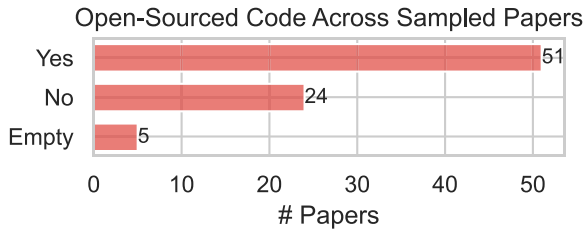

Figure 6: Proportion of papers with open-source repositories from our sample of peer-reviewed papers.

that work in this domain is justified by the need to ‘‘raise awareness’’ of newly uncovered security vulnerabilities (e.g., Yang et al., 2021b; Qi et al., 2021b,c; Chen et al., 2022b). If we accept the norm from cybersecurity—that a full public disclosure should typically come with the necessary code to assist other ethical researchers and practitioners—NLPSec falls short of this ideal. From our sample of works, 36% are functionally closed-source (see Figure 6). As a consequence, **white hat NLPSec researchers and practitioners may be at a disadvantage when it comes to securing systems**. Depending on the severity of a vulnerability discussed in a given work, malicious actors may be able to benefit from the latency of white hat engineers re-implementing a paper’s threat model. The relationship between public disclosure and harm minimization is explored in Section 5.1.

5 Discussion

While NLPSec is an interdisciplinary field with an indisputable connection to cybersecurity, there are cases where the parallels between these fields diverge. In this section, we explore some areas where the analogy between NLPSec and cybersecurity fails, with the goal of illuminating the urgent need for a broader conversation about ethics in NLPSec. To help initiate such a conversation, we conclude with some concrete recommendations for NLPSec researchers, towards adopting better practices for more ethical NLPSec research.

5.1 To Name It Is To Own It: Misuse and Other Harms

In our survey, *misuse* was the most commonly cited potential harm inherent to research in NLPSec (37% of works in Figure 4). The threat of misuse can be understood to result from public disclosure, as malicious actors are known

weaponize information therein (Kokkinakis et al., 2022). With the goal of preventing misuse, one reactionary approach would be to resign ourselves from publicly sharing such sensitive work (i.e., *security through obscurity* (Guo et al., 2018)). This approach has been largely rejected by the wider cybersecurity community, as security through obscurity is difficult to maintain, creates a false sense of security, and clashes with the scientific value of transparency (e.g., Courtois, 2009). Works in NLPSec are thus caught in what has been dubbed *The Devil's Triangle* (Thieltges et al., 2016; Leidner and Plachouras, 2017): The path towards model security hinges upon transparency, which is required for researchers to make progress, but is also advantageous for cybercriminals, while innocent actors can be harmed in the cross-fire. To help NLPSec escape the Devil's Triangle, we look to existing works in NLP on misuse, and examine whether these suggestions make sense in NLPSec.

Dual Use and Misuse in NLP NLP technologies like LLMs can be leveraged for a wide variety of applications, ranging from the virtuous (e.g., improving accessibility), to the reprehensible (e.g., proliferating hate speech) and the innocuous (e.g., generating fan fiction). That these models can simultaneously be utilized for *both* ''legitimate'' and ''illegitimate'' purposes is commonly referred to as **dual use** (Riecke, 2023). Traditionally, dual use has been viewed primarily through the lens of a ''civilian'' versus ''military'' dichotomy in terms of applications, but due to the mass availability of NLP tools, there are also opportunities for civilian black hats to use the technology in unsavory ways outside the scope of warfare. In the context of NLP, dual use has only recently been discussed in depth by Kaffee et al. (2023). In their work, they define dual use in NLP as ''malicious reuse'' (i.e., **misuse**, where the intended purpose of the technology is violated). Examples of misuse across NLP include: manipulating models for automated influence operations (e.g., misinformation) (Goldstein et al., 2023), surveillance of marginalized groups (Sannon and Forte, 2022), and by-passing safety features to generate hate speech or otherwise engage in illegal activities (e.g., phishing) (Yong et al., 2024). As black hat hackers are known to misuse white hat software (Martin, 2017), the threat of misuse against NLPSec research is palpable.

To this end, Kaffee et al.'s (2023) work focuses on traditional NLP, and thus the scope of their exploration is not configured to accommodate the unique position of NLPSec, as an interdisciplinary field. While the authors briefly mention Henderson et al. (2023) (who propose a defense method for preventing malicious use-cases of LLMs), as well as the possible criminal applications of LLMs for phishing, Kaffee et al.'s (2023) proposed checklist does not help NLPSec practitioners to better navigate the problems of misuse. For example, their checklist for preventing misuse includes the following questions:

1. Can any scientific artifacts you create be used for military[6] application?
2. Can any scientific artifacts you create be used to harm or oppress any and particularly marginalised groups of society?
3. Can any scientific artifacts you create be used to intentionally manipulate, such as spread disinformation or polarize people?

In the context of NLPSec, the answer to the above questions will typically be ''yes''. Additionally, we observe that most papers only consider dual use or misuse as an afterthought, if at all, in line with the results of Kaffee et al. (2023). This concerning trend further stresses the importance of a discussion on misuse, tailored to NLPSec.

Other Potential Harms in NLPSec The findings of our survey indicate that practitioners in NLPSec widely disagree about what potential harms are inherent to this research, making it difficult to ask NLPSec to blanketly minimize harm. Part of this nonagreement may stem from a historical understanding of how harm minimization is typically discussed in NLP. Traditionally in NLP, practices of harm minimization have concerned very different issues than cybersecurity. Where cybersecurity is concerned with criminality, NLP has historically focused on fair payment of crowd-workers (Shmueli et al., 2021), and the prioritization of researching techniques that

[6]Military funding has a long and complicated history in the sciences (Smit, 1995). While we do not examine the sources of funding across our sample of papers, we note the presence of institutions associated with the military and defense industry among the author affiliations.

directly combat known flaws of LLMs (Weidinger et al., 2021), like dissemination of harmful social biases (Brown et al., 2020; Abid et al., 2021; Lucy and Bamman, 2021) and misinformation (Lewis and Marwick, 2017; Kenton et al., 2021). As discussed by Leidner and Plachouras (2017), it is important that NLP researchers proactively plan to avoid unethical scenarios. As NLPSec combines the potential harms of NLP with those of cybersecurity, the need to anticipate and mitigate risks is crucial.

To promote both effective and ethically responsible NLPSec research, we emphasize minimizing harm as a *design principle*. Specifically, conversations about harm minimization should take place throughout a given project's research life-cycle, from the initial planning, funding, and designing of a project, to publishing and disseminating the conclusions (Galinkin, 2022). This process ensures that research ethics remain core considerations throughout the work, rather than a mere rhetorical post-hoc ethical statement (Peters et al., 2020). Additionally, Gardner et al. (2022) emphasize the role of funding agencies in ensuring trustworthy AI, by *mandating* ethical assessments throughout the application, evaluation, and implementation phases, both from applicants and the funding agencies, aided by experts in ethics. Presently, such strict ethical requirements are not the norm in NLP, however. Until there are strong ethics review requirements across the field, as with other sciences, it is imperative that researchers clearly articulate the potential for dual use and misuse in their works, and provide viable defenses for the most vulnerable scenarios.

In NLPSec, positive examples of harm minimization have prioritized user safety, avoiding scenarios which could expose sensitive data or scenarios that directly affect end-users. For example, Parikh et al.'s (2022) methodology intentionally avoids exposing sensitive data of real-world users in a data reconstruction attack by planting synthetic ''canaries'' (i.e., fake instances of private data) into their training data. Similarly, in their work exploring adversarial attacks, Song et al. (2021) do *not* experiment with real-world systems, such that no end users are harmed.

5.2 The Victims of English-Centric NLPSec

Below, we explore the role of multilinguality in NLPSec, the urgency of unresolved security vulnerabilities in relation to lower-resourced languages, and how traditional norms of consent in NLP may conflict with those of cybsersecurity.

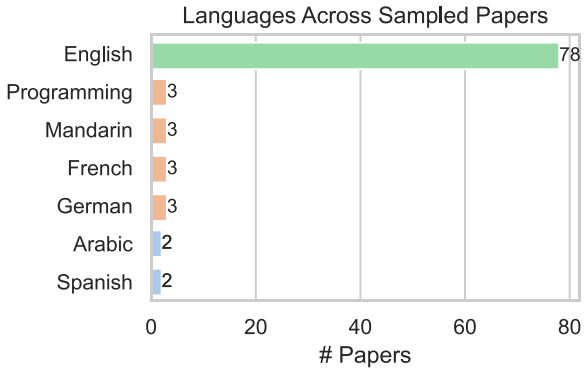

Figure 7: Of the victim languages investigated across the sample of 80 papers, we identify 22 *natural* languages. Languages not displayed above (as n = 1) are Japanese (Zeng and Xiong, 2021), Tibetan (Cao et al., 2023), Javanese, Indonesian, Malaysian, Tagalog, Tamil (Wang et al., 2024b), and the remaining languages of the XNLI dataset: Greek, Bulgarian, Russian, Turkish, Vietnamese, Thai, Hindi, Swahili, and Urdu (Lin et al., 2024).

Security As Strong As Its Weakest Link As one of the most well represented languages in NLP, it is no surprise that English is present in 97% of the works sampled (see Figure 7). Emerging research examining multilingual NLPSec, however, suggests that that multilingual models may be *more* vulnerable to attacks than their monolingual (English) counterparts, as demonstrated in the context of embedding inversion attacks (Chen et al., 2024b, 2025) and backdoor attacks (He et al., 2024). Additionally, recent works also show how lower-resourced languages can be weaponized to bypass LLM safety features (Yong et al., 2024), as well as introduce backdoors (Wang et al., 2024b), creating further cause for concern. To this end, Yong et al. (2024) discuss the apparent shift in consequences for poor performance over lower-resourced languages: previously, a lack of competitive models to handle these languages culminated primarily in technological disparity, affecting only the community in question. This inequality can be exploited by malicious actors, resulting in a threat to everyone. In other words, the security of NLP models is now only as strong as its weakest link. Alarmingly, such weaponization of under-performing language technologies is already being observed (Nigatu and Raji, 2024).

Higher Stakes For Lower Resourced Scenarios Low-resource languages can often be situated in vulnerable contexts, which brings heightened need to protect the communities speaking them. For example, previous works in NLP have exposed correlations between GDP and data availability (Blasi et al., 2022; Ranathunga and de Silva, 2022), underscoring how the gap between higher- and lower-resource languages is part of a broader picture of global inequality. At the same time, even within wealthier nations, minority languages (often low-resource) may require revitalization efforts in order to stave off extinction (e.g., Indigenous languages of Australia (Meakins and O'Shannessy, 2016)); and other widely adopted languages may battle stigma, obstructing their inclusion in language technology (e.g., Creoles (Lent et al., 2024)). Practitioners in low-resource NLP have developed their own ethical norms in response to such concerns, grounded in the prioritization of a community's specific needs and aspirations for language technology, as well as the preservation of their autonomy (Bird, 2022; Lent et al., 2022; Mager et al., 2023).

Research Traditions Collide: Consent Consent is highly context-dependent in both cybersecurity and general NLP research ethics, leading to potential conflicts when these fields intersect in NLPSec. In traditional cybersecurity, the necessity of consent largely depends on the nature of the system being tested. When testing systems running on third-party infrastructure, such as web services or cloud platforms, obtaining explicit consent is expected in order to, e.g., avoid legal or ethical issues. However, consent is not typically required for, e.g., research involving hardware or software running locally on the researcher's infrastructure, even if it violates end-user license agreements (EULAs) or terms-of-service (ToS) contracts (Kozhuharova et al., 2022). For example, reverse engineering or probing locally deployed systems for vulnerabilities is widely accepted as a valid and necessary practice, provided it prioritizes public safety and minimizes harm.

In contrast, multilingual NLP research has increasingly emphasized *community* consent, particularly regarding low-resource languages or marginalized groups. These efforts are grounded in the principle of respecting the autonomy and cultural context of the communities whose languages and data are being studied. NLPSec introduces scenarios where these norms may conflict. For example, securing LLMs for low-resource languages is vital for ensuring downstream safety of their communities. However, using a language for security testing without explicit consent risks reducing it to a mere tool for experimentation, potentially alienating the communities involved and exacerbating existing inequalities (Bird, 2020).

This aspect is largely ignored in NLPSec to date. Among our sample, only one work engaged with a truly low-resource language (see Figure 7). In this study, Cao et al. (2023) aim to raise awareness about the threatened security of minority languages by contributing a script-based adversarial attack method for Tibetan, a notably vulnerable language[7] against CINO (Yang et al., 2022).[8]

Ultimately, the ''white hat'' versus ''black hat'' paradigm in cybersecurity is further complicated when extended to NLPSec. While cybersecurity often frames consent as an ethical trade-off, NLPSec researchers must also mind the long-standing ethical norms of NLP, particularly for lower-resourced or otherwise marginalized languages. Balancing these differing research norms around consent demands careful consideration. On the one hand, engaging with language community representatives and aligning research with their needs can ensure that low-resource languages are not exploited in ways that harm or undermine their speakers. On the other hand, delaying research to secure consent could leave vulnerable languages at greater risk of exploitation by malicious actors. Given the present and severe threat of weaponization of lower-resourced languages, however, this issue must not remain unaddressed by NLPSec; prioritization of ethical research practices will be critical for harm minimization.

5.3 Obstacles for CVD in NLPSec

The results of our survey showed that no works *report* whether efforts to do CVD (see Section 2.2) occurred. Outside the scope of our survey, positive examples of CVD in NLPSec do exist, such as Carlini et al. (2024), who state clearly in their

[7]Tibetan has relatively few speakers (1.2 million native speakers according to `https://en.wikipedia.org/wiki/Lhasa_Tibetan` and 6 million speakers in total (Tournadre, 2013)), is actively undergoing revitalization (Roche and Bum, 2018), and is situated in a delicate sociopolitical context (Roche, 2017; Jia and Qie, 2021).

[8]CINO a pre-trained multilingual LLM for handling languages spoken across China (i.e., Mandarin, Cantonese, Korean, Mongolian, Uyghur, Kazakh, Zhuang, and Tibetan).

manuscript that the discovered vulnerability was reported to OpenAI and that the release of the paper followed the company's response to and mitigation of the risk described in their work. In this section, we aim to explore some reasons why CVD is a potentially nontrivial in NLPSec.

Not Fixable by One Line of Code Models are one part of complex computer software systems subject to more general cybersecurity vulnerabilities, e.g., remote code execution, as has been observed in the `llama-cpp-python` library.[9] In such instances, CVD offers NLPSec researchers a way to maintain transparency, without assisting malicious actors, as the published vulnerabilities will hopefully have already been patched. In NLP and the broader machine learning space, however, this remediation and mitigation process is complicated by the fact that **discovered issues may be endemic** to the target of evaluation or require prohibitively expensive retraining to fix. As opposed to traditional software, one cannot simply write a patch that fixes a discovered issue entirely, and instead, guidance must be provided to users in order to allow them to accept or mitigate risk appropriately. In one recent example, the LAION 5B dataset was found to contain child sex abuse material (Birhane et al., 2021, 2023). The dataset was accordingly taken offline by its authors (LAION.ai, 2023; Thiel, 2023). Consequently, all models trained on this dataset were at known risk of producing illegal, harmful materials. Model providers, upon being made aware of this risk, had the obligation to decide whether to retrain the model or accept the risk—this was not something that could be managed by updating a few lines of code. Still, disclosure of the risk allows affected parties to make informed decisions.

An Open Problem for Open Models While CVD is most relevant for research using proprietary models, it also remains relevant for open-weight, freely available models. As the majority of works in NLPSec have thus far been concerned with attacking or defending open-weight models (see Figure 5), CVD may seem less applicable to models that are not actively maintained by the organization hosting them. In such a case where the practical steps towards CVD may be unclear, an open question in NLPSec is how to best disclose risks—if at all—both to the pertinent organizations and to the broader scope of users. Similar to the trolley problem introduced by Kohno et al. (2023) (Section 2.2), there may be instances where CVD requires careful consideration, for example in critical sectors like healthcare. However in the majority of NLP, where a model can typically be replaced with another with relative ease, it is difficult to give a generalized conjecture on such cases. For our part, we recommend that NLP researchers engage in best-effort attempts to alert model providers to potential risks. Most companies that provide models as a service have a channel for external users to file bug reports. In cases where the model is produced by a smaller entity, opening issues on the platform where the model is shared, e.g., Github, HuggingFace, or emailing authors of a paper tied to the model serves as a good channel for attempting this coordination before publication of results.

Another complication for CVD in NLPSec is scalability. As a field, NLP places immense value on scalability (Kogkalidis and Chatzikyriakidis, 2024). Researchers are often expected or encouraged to massively scale their experiments to an increasing number of models and languages. While scaling experiments largely hinges upon the availability of compute resources, the process of CVD does not scale so. Ideally, CVD entails intentional, personal communication between the researcher and the affected organization. Responsible disclosure to CERTs or other trusted communities functions the same way. The human aspect of this process cannot simply be outsourced and automated to a machine, thus conflicting with the expectations for massive scalability within NLPSec.

5.4 Recommendations for NLPSec Practitioners

Thusfar, the field of NLPSec has largely been operating in a ''gray hat'' manner, where the individual researcher is compelled to rely on their own moral compass. This is in part due to the overwhelming bulk of AI regulatory documents (Larsson, 2021), which often do not directly relate to a security angle, as well as the rigidity of applying certain cyber security ethical norms to the unique problems of NLPSec. In response to this pressure point, we aim to provide some concrete

[9] `https://github.com/abetlen/llama-cpp-python/security/advisories/GHSA-56xg-wfcc-g829`.

recommendations to help the field take concrete steps towards more ethical NLPSec. To this end, we hope future works will benefit from, and build upon, the following recommendations:

1. **Plan Ahead to Minimize Harm:** Ethical considerations should not be relegated to a post-hoc ethics statement. First, consider the harms entailed in conducting a study and in foregoing it. Design experiments with harm minimization in mind from the start. Include these details in the main body of your work. Beyond reducing the potential harms of research and helping researchers avoid downstream ethical conundrums, this approach also promotes a culture of responsible research.

2. **Prioritize Multilingual Equity:** Include multilingual models and lower-resourced languages in NLPSec work to build towards comprehensive security coverage for all. Prioritize typologically diverse language samples (Ploeger et al., 2024). Engage with the communities speaking these languages to seek consent and avoid exploitation. Consider whether a particular community might be jeopardized as a result of your work. Researchers should respect the autonomy of these communities, while working to address the established heightened vulnerabilities in such low-resource scenarios.

3. **Approach Disclosure Responsibly: (a)** Consider the most appropriate options for disclosure. If you can complete CVD, contact relevant parties about security breaches 60–90 days prior to any publication and clearly acknowledge that CVD occurred directly in the published manuscript. Even for open-weight models, best-effort attempts to alert stakeholders should be made, such as model providers or the platforms hosting the models. If you cannot complete CVD, attempt responsible disclosure to other affected parties. Decide whether public disclosure is appropriate, and act accordingly. Communicate this thought process in your paper. **(b)** When appropriate, release accompanying proof-of-concept code to help NLPSec researchers better defend against attacks. While black hats will always make time to re-implement attacks for nefarious gains, white hats are time-constrained and defense becomes harder without clear & explicit resources. If not appropriate, explain why in your manuscript. Ask yourself if public disclosure is still warranted, if open-sourcing code is not.

6 Conclusion

In the burgeoning field of NLPSec, most works consider scenarios where a malicious attacker seeks to undermine a system's intended behavior, with the goal of causing harm. Research output in NLPSec thus stands to be highly consequential in the face of mass-adoption of language technologies such as LLMs, and its relevance to public safety necessitates heightened scrutiny when it comes to best practices for ethical research. Given NLPSec's position as a truly interdisciplinary field, practitioners in this space can benefit from the rich traditions of research ethics from both cybersecurity and NLP. In this work, however, we find that NLPSec works published in NLP venues generally fall short of the ethical standards set by cybersecurity (Section 4), signaling a higher-level disconnect between NLP and cybersecurity practitioners for work in this area. This failure to inherit ethical best practices can arise from a variety factors (Section 5), but largely stem from the differences between traditional cybersecurity and NLPSec, which underscores the limitations of the ''white versus black hat'' paradigm of cybersecurity as applied to NLPSec. Still, we argue that the repercussions of the current research patterns are grim: works in NLPSec may benefit would-be attackers more than the public (Section 2), with dire consequences for everyone, but especially for already-marginalized communities (Section 5.2). By highlighting these problems and exploring their nuances, this work aims to persuade the field of the urgency of the present situation and to spark a much-needed conversation across the field of NLPSec. To kick off this conversation, we provide some concrete recommendations to help practitioners transition from gray hat to white hat NLP.

Limitations

Defining Ethical Hacking While ethical hacking is, on its face, a noble venture, it is **a term that features some subjectivity and may find itself**

at odds with the desires of particular groups or individuals. For instance, the definition of an ethical hacker as one who is ''trustworthy for business and lawful'' (Christen et al., 2020) may run headlong into both trust and the law. Regarding trust, many organizations have adopted an approach that is far more friendly to security researchers, but there are organizations who are notorious to this day for their attempts to keep the discovery of vulnerabilities in their products quiet. As it concerns the law, there are two primary issues to contend with. First, what is ''lawful'' will necessarily change across jurisdictions, with laws differing not merely between countries, but sometimes across provinces and states. For example, the ethics of government-associated cybersecurity research (e.g., government hacking) can be a topic of debate, even when practitioners are acting under the color of law. Another example includes the exemption for security research in the United States, which is not written into law, but is rather part of the US Department of Justice's prosecution guidelines (U.S. Department of Justice, 2022), updated in 2022, indicating that good faith security research should not be prosecuted. In other words, much ethical hacking in the US may be considered unlawful but will simply not be prosecuted. The second issue is time. Across jurisdictions, laws are likewise positioned to evolve over time, especially as the list of known cyber-threats grows to include attacks against NLP models. In general, as it is often difficult for the law to keep up with rapid technological progress, ethics training must be prioritized in both the NLP (Bender et al., 2020) and cybersecurity curriculum (Blanken-Webb et al., 2018).

Risks of Monocultural Ethics Discourse As AI becomes further entrenched in daily life, the risks imposed from research and commercial activities in AI are also a global issue. Similar to how correspondents of Kaffee et al.'s (2023) survey are overwhelmingly from a Western audience,[10] the bulk of AI governance documents are also overwhelmingly of Western origin (e.g., Larsson, 2021; UNESCO, 2021). In contrast, Figure 10 (Appendix B) reveals that the majority of NLPSec research comes from Asia. However, historical and cultural differences have led to fundamentally different approaches to addressing AI risks across these regions.[11] For example, while the deployment of technologies such as facial recognition is illegal and considered strictly unethical in the EU because of GDPR (European Parliament and Council of the European Union, 2016), it is widely deployed in countries such as China (Dudley, 2020), Iran (George, 2023), Canada (CCLA, 2001), and the US (GAO, 2023), where the local personal data is collected, raising concerns over human rights. Still, models trained on such data may be imported to the EU and deployed without any legal consequences, highlighting a global ethical risk, termed *ethics dumping* (Commission et al., 2013; ECDGRI, 2016), where non-ethical practices are shifted to countries lacking certain ethics regulations. This divide between AI governance and the regions impacted by AI calls for inclusion of diverse perspectives, especially in a burgeoning and cross-disciplinary field like NLPSec. Of course, cross-cultural AI ethics is notably diverse. For example, African Ubuntu philosophy promotes communal values in the use of AI (Gwagwa et al., 2022), Abrahamic religious views stress that AI use should respect human dignity (Goltz et al., 2020; Raquib et al., 2022), and Buddhist AI philosophy advocates for reducing pain and suffering using AI (Hughes, 2012; Hongladarom, 2021). When faced with seemingly irresolvable conflicts of ethical values across cultures, we urge NLPSec researchers to look towards the UN Declaration of Human Rights, which outlines the fundamental rights and freedoms of all human beings.

Ethics Statement

Our work adheres to the ACM Code of Ethics. As we analyze, e.g., author metadata, we have ensured that the licenses of the data sources allow for this type of data extraction. This line of work, including our methodology and the analysis of author meta-data, has received approval from the Aalborg University Research Ethics Committee under case number 2024-505-00376.

Acknowledgments

H. L., Y. C., and J. B. are funded by the Carlsberg Foundation, under the Semper Ardens: Accelerate programme (project nr. CF21-0454). This work

[10] Of 48 participants, only 3 hailed from Asia and 1 from Africa, with the remainder from Europe or North America.

[11] The authors acknowledge that we represent Western institutions and have our own values and biases accordingly.

benefited greatly from help and conversations with others. Thank you to Christopher Fiorelli and Maria Antoniak at AI2 for their help with the Semantic Scholar crawler, to Zeerak Talat for detailed feedback on our manuscript, and to Steven Bird for conversations on ethics of AI, which enriched this manuscript. Thank you to members of the AAU NLP Research Group for their feedback during paper clinics, to Mike Zhang for help improving figures, to Nicholas Walker for feedback on multiple drafts of this paper, and to Shreyas Srinivasa for initial input on cybersecurity research norms. Finally, thank you to the TACL area chair and reviewers, whose constructive feedback played a major role in shaping this manuscript.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models.

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Made laine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Jo hannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such,

Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe, Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Jennifer Ackerman. 2022. Regina couple says possible ai voice scam nearly cost them $9,400. *Regina Leader Post*.

Norah Alshahrani, Saied Alshahrani, Esma Wali, and Jeanna Matthews. 2024. Arabic synonym BERT-based adversarial examples for text classification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 137–147, St. Julian's, Malta. Association for Computational Linguistics.

Michael Anderson and Susan Leigh Anderson. 2007. Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4):15–15.

Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. 2012. The menlo report. *IEEE Security and Privacy*, 10(2):71–75. https://doi.org/10.1109/MSP.2012.52

Rongzhou Bao, Jiayi Wang, and Hai Zhao. 2021. Defending pre-trained language models from adversarial word substitution without performance sacrifice. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3248–3258, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.287

Emily M. Bender, Dirk Hovy, and Alexandra Schofield. 2020. Integrating ethics into the NLP curriculum. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-tutorials.2

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.313

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.539

Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. 2023. Into the laions den: Investigating hate in multimodal datasets.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: Misogyny, pornography, and malignant stereotypes.

Jane Blanken-Webb, Imani Palmer, Nicholas C. Burbules, Roy H. Campbell, and Masooda N. Bashir. 2018. A case study-based cybersecurity ethics curriculum. In *ASE @ USENIX Security Symposium*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.376

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter,

Christopher 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.

Aaron J. Burstein. 2008. Conducting cybersecurity research legally and ethically. *LEET*, 8:1–8.

Xi Cao, Dolma Dawa, Nuo Qun, and Trashi Nyima. 2023. Pay attention to the robustness of Chinese minority language models! Syllable-level textual adversarial attack on Tibetan script. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 35–46, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.trustnlp-1.4

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick, and Florian Tramèr. 2024. Stealing part of a production language model.

Canadian Civil Liberties Association CCLA. 2001. Police use of facial recognition technology in canada and the way forward.

Shuguang Chen, Leonardo Neves, and Thamar Solorio. 2024a. Context-aware adversarial attack on named entity recognition. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 11–16, San Ġiljan, Malta. Association for Computational Linguistics.

Sishuo Chen, Wenkai Yang, Zhiyuan Zhang, Xiaohan Bi, and Xu Sun. 2022a. Expose backdoors on the way: A feature-based efficient defense against textual backdoor attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 668–683, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.47

Yiyi Chen, Russa Biswas, Heather Lent, and Johannes Bjerva. 2025. Against all odds: Overcoming typology, script, and language confusion in multilingual embedding inversion attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, United States. AAAI Press. https://doi.org/10.1609/aaai.v39i22.34533

Yiyi Chen, Heather Lent, and Johannes Bjerva. 2024b. Text embedding inversion security for multilingual language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7808–7827, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.422

Yangyi Chen, Fanchao Qi, Hongcheng Gao, Zhiyuan Liu, and Maosong Sun. 2022b. Textual backdoor attacks can be more harmful via two simple tricks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11215–11221, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.770

Yangyi Chen, Jin Su, and Wei Wei. 2021. Multi-granularity textual adversarial attack with behavior cloning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4511–4526, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.371

YunSeok Choi, Hyojun Kim, and Jee-Hyong Lee. 2022. TABS: Efficient textual adversarial attack for pre-trained NL code model using semantic beam search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5490–5498, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.369

M. Christen, B. Gordijn, and M. Loi. 2020. *The Ethics of Cybersecurity*. The International Library of Ethics, Law and Technology. Springer International Publishing. https://doi.org/10.1007/978-3-030-29053-5

EC-European Commission et al. 2013. Horizon 2020 work programme 2014–2015. *Science with and for Society*.

Nicolas T. Courtois. 2009. The dark side of security by obscurity and cloning MiFare classic rail and building passes anywhere, anytime. Cryptology ePrint Archive, Paper 2009/137. https://eprint.iacr.org/2009/137

Daniel C. Dennett. 1997. When hal kills, who's to blame? Computer ethics. *HAL's Legacy: 2001's Computer as Dream and Reality*, pages 351–365.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Du, Peixuan Li, Haodong Zhao, Tianjie Ju, Ge Ren, and Gongshen Liu. 2024. UOR: Universal backdoor attacks on pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7865–7877, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.1287/opre.2021.0562

Lauren Dudley. 2020. China's ubiquitous facial recognition tech sparks privacy backlash.

European Commission Directorate-General for Research Innovation ECDGRI. 2016. H2020 programme guidance: How to complete your ethics self-assessment.

Adel Elmahdy and Ahmed Salem. 2024. Deconstructing classifiers: Towards a data reconstruction attack against text classification models. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 143–158, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.privatenlp-1.15

European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council.

Courtney Falk. 2004. Gray hat hacking: Morally black and white.

Xuanjie Fang, Sijie Cheng, Yang Liu, and Wei Wang. 2023. Modeling adversarial attack on pre-trained language models as sequential decision making. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7322–7336, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.461

Erick Galinkin. 2022. Towards a responsible AI development lifecycle: Lessons from information security. *ArXiv*, abs/2203.02958.

Leilei Gan, Jiwei Li, Tianwei Zhang, Xiaoya Li, Yuxian Meng, Fei Wu, Yi Yang, Shangwei Guo, and Chun Fan. 2022. Triggerless backdoor attack for NLP tasks with clean labels. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2942–2952, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.214

Chongyang Gao, Kang Gu, Soroush Vosoughi, and Shagufta Mehnaz. 2024. Semantic-preserving adversarial example attack against BERT. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 202–207, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.trustnlp-1.17

US Government Accountability Office GAO. 2023. Facial recognition services: Federal law enforcement agencies should take actions to implement training, and policies for civil liberties.

Allison Gardner, Adam Leon Smith, Adam Steventon, Ellen Coughlan, and Marie Oldfield. 2022. Ethical funding for trustworthy AI: Proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice. *AI and Ethics*, pages 1–15. https://doi.org/10.1007/s43681-021-00069-w, PubMed: 34790951

Rachel George. 2023. The AI assault on women: What iran's tech enabled morality laws indicate for women's rights movements.

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations.

Nachshon (Sean) Goltz, John Zeleznikow, and Tracey Dowdeswell. 2020. From the Tree of Knowledge and the Golem of Prague to Kosher autonomous cars: The ethics of artificial intelligence through Jewish eyes. *Oxford Journal of Law and Religion*, 9(1):132–156. https://doi.org/10.1093/ojlr/rwaa015

Victoria Graf, Qin Liu, and Muhao Chen. 2024. Two heads are better than one: Nested PoE for robust defense against multi-backdoors. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 706–718, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.40

Wenbo Guo, Qinglong Wang, Kaixuan Zhang, Alexander G. Ororbia, Sui Huang, Xue Liu, C. Lee Giles, Lin Lin, and Xinyu Xing. 2018. Defending against adversarial samples without security through obscurity. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 137–146. IEEE. https://doi.org/10.1109/ICDM.2018.00029

Arthur Gwagwa, Emre Kazim, and Airlie Hilliard. 2022. The role of the african value of ubuntu in global AI inclusion discourse: A normative ethics perspective. *Patterns*, 3(4). https://doi.org/10.1016/j.patter.2022.100462, PubMed: 35465235

Wenjuan Han, Liwen Zhang, Yong Jiang, and Kewei Tu. 2020. Adversarial attack and defense of structured prediction models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2327–2338, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.182

Ishrak Hayet, Zijun Yao, and Bo Luo. 2022. Invernet: An inversion attack framework to infer fine-tuning datasets through word embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5009–5018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.368

Julian Hazell. 2023. Spear phishing with large language models. *arXiv preprint arXiv:2305.06972*.

Xuanli He, Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2023a. IMBERT: Making BERT immune to insertion-based backdoor attacks. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 287–301, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.trustnlp-1.25

Xuanli He, Jun Wang, Qiongkai Xu, Pasquale Minervini, Pontus Stenetorp, Benjamin I. P. Rubinstein, and Trevor Cohn. 2024. Transferring troubles: Cross-lingual transferability of backdoor attacks in LLMs with instruction tuning.

Xuanli He, Qiongkai Xu, Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2023b. Mitigating backdoor poisoning attacks through the lens of spurious correlation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 953–967, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.60

Peter Henderson, Eric Mitchell, Christopher Manning, Dan Jurafsky, and Chelsea Finn. 2023. Self-destructing models: Increasing the costs of harmful dual uses of foundation models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 287–296, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3600211.3604690

Kenneth Einar Himma and Herman T. Tavani. 2008. *The handbook of information and computer ethics*. Wiley Online Library.

Soraj Hongladarom. 2021. What buddhism can do for AI ethics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2096

Hai Huang, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. 2024a. Composite backdoor attacks against large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1459–1472, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.94

Yangsibo Huang, Zhao Song, Danqi Chen, Kai Li, and Sanjeev Arora. 2020. TextHide: Tackling data privacy in language understanding tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1368–1382, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.123

Yu-Hsiang Huang, Yuche Tsai, Hsiang Hsiao, Hong-Yi Lin, and Shou-De Lin. 2024b. Transferable embedding inversion attack: Uncovering privacy risks in text embeddings without model queries. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4193–4205, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.230

James Hughes. 2012. Compassionate AI and selfless robots: A buddhist approach. *Robot ethics: The ethical and social implications of robotics*, pages 69–83.

Nanna Inie, Jonathan Stray, and Leon Derczynski. 2025. Summon a demon and bind it: A grounded theory of LLM red teaming. *PLOS One*, 20(1):e0314658. https://doi.org/10.1371/journal.pone.0314658, PubMed: 39813184

ISO 29147:2018. 2018. Information technology — security techniques — vulnerability disclosure. Standard, International Organization for Standardization, Geneva, CH.

Luo Jia and Pai Qie. 2021. A sociological analysis of tibetan language policy issues in china. *SN Social Sciences*, 1:1–31. https://doi.org/10.1007/s43545-021-00092-y

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Lesheng Jin, Zihan Wang, and Jingbo Shang. 2022. WeDef: Weakly supervised backdoor defense for text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11614–11626, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.798

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, pages 1–11. https://doi.org/10.1038/s42256-019-0088-2

Lucie-Aimée Kaffee, Arnav Arora, Zeerak Talat, and Isabelle Augenstein. 2023. Thorny roses: Investigating the dual use dilemma in natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13977–13998, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.932

Yannik Keller, Jan Mackensen, and Steffen Eger. 2021. BERT-defense: A probabilistic model based on BERT to combat cognitively inspired orthographic adversarial attacks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1616–1629, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.141

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents.

Donggyu Kim, Garam Lee, and Sungwoo Oh. 2022. Toward privacy-preserving text embedding similarity with homomorphic encryption.

In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 25–36, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.finnlp-1.4

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F. Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R. Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D. Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Konstantinos Kogkalidis and Stergios Chatzikyriakidis. 2024. On tables with numbers, with numbers.

Tadayoshi Kohno, Yasemin Acar, and Wulf Loh. 2023. Ethical frameworks and computer security trolley problems: Foundations for conversations. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5145–5162, Anaheim, CA. USENIX Association.

Dimitrios Kokkinakis, Charalambos K. Themistocleous, Kristina Lundholm Fors, Athanasios Tsanas, and Kathleen C. Fraser, editors. 2022. *Proceedings of the RaPID Workshop - Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments - within the 13th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France.

Denitsa Kozhuharova, Atanas Kirov, and Zhanin Al-Shargabi. 2022. Ethics in cybersecurity. What are the challenges we need to be aware of and how to handle them? In *Cybersecurity of Digital Service Chains: Challenges, Methodologies, and Tools*, pages 202–221. Springer International Publishing Cham. https://doi.org/10.1007/978-3-031-04036-8_9

LAION.ai. 2023. Safety review for laion-5b. https://laion.ai/notes/laion-maintenance/. Accessed: 2024-03-05.

Stefan Larsson. 2021. AI in the EU: Ethical guidelines as a governance tool. *The European Union and the Technology Shift*, pages 85–111. https://doi.org/10.1007/978-3-030-63672-2_4

Yibin Lei, Yu Cao, Dianqi Li, Tianyi Zhou, Meng Fang, and Mykola Pechenizkiy. 2022. Phrase-level textual adversarial attack with label preservation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1095–1112, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-naacl.83

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1604

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders Søgaard, and Johannes Bjerva. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of*

the Association for Computational Linguistics, 12:950–978. https://doi.org/10.1162/tacl_a_00682

Becca Lewis and Alice E. Marwick. 2017. Media manipulation and disinformation online.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021a. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online. Association for Computational Linguistics.

Haoran Li, Mingshi Xu, and Yangqiu Song. 2023a. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14022–14040, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.881

Jiazhao Li, Zhuofeng Wu, Wei Ping, Chaowei Xiao, and V. G. Vinod Vydiswaran. 2023b. Defending against insertion-based textual backdoor attacks via attribution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8818–8833, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.561

Jiazhao Li, Yijin Yang, Zhuofeng Wu, V. G. Vinod Vydiswaran, and Chaowei Xiao. 2024. ChatGPT as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2985–3004, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.165

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.500

Linyang Li, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu. 2021b. Backdoor attacks on pre-trained models by layerwise weight poisoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3023–3032, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.241

Linyang Li, Demin Song, and Xipeng Qiu. 2023c. Text adversarial purification as defense against adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 338–350, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.20

Yanzhou Li, Shangqing Liu, Kangjie Chen, Xiaofei Xie, Tianwei Zhang, and Yang Liu. 2023d. Multi-target backdoor attacks for code pre-trained models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7236–7254, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.399

Zichao Li, Dheeraj Mekala, Chengyu Dong, and Jingbo Shang. 2021c. BFClass: A backdoor-free text classification framework. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 444–453, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.40

Yu Lin, Qizhi Zhang, Quanwei Cai, Jue Hong, Wu Ye, Huiqi Liu, and Bing Duan. 2024. An inversion attack against obfuscated embedding matrix in language model inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2100–2104, Miami, Florida, USA. Association for Computational

Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.126

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Zhengxiao Liu, Bowen Shen, Zheng Lin, Fali Wang, and Weiping Wang. 2023. Maximum entropy loss, the silver bullet targeting backdoor attacks in pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3850–3868, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.237

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nuse-1.5

Weimin Lyu, Songzhu Zheng, Tengfei Ma, and Chao Chen. 2022. A study of the attention abnormality in trojaned BERTs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4727–4741, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.348

Kevin Macnish and Jeroen Van der Ham. 2020. Ethics in cybersecurity research and practice. *Technology in Society*, 63:101382. https://doi.org/10.1016/j.techsoc.2020.101382

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.268

C. Dianne Martin. 2017. Taking the high road white hat, black hat: The ethics of cybersecurity. *ACM Inroads*, 8(1):33–35. https://doi.org/10.1145/3043955

Andrea M. Matwyshyn, Ang Cui, Angelos D. Keromytis, and Salvatore J. Stolfo. 2010. Ethics in security vulnerability research. *IEEE Security & Privacy*, 8(2):67–72. https://doi.org/10.1109/MSP.2010.67

Felicity Meakins and Carmel O'Shannessy. 2016. *Loss and Renewal: Australian Languages since Colonisation*, volume 13. Walter de Gruyter GmbH & Co KG. https://doi.org/10.1515/9781614518792

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2023. Tree of attacks: Jailbreaking black-box llms automatically. *arXiv preprint arXiv:2312.02119*.

Kai Mei, Zheng Li, Zhenting Wang, Yang Zhang, and Shiqing Ma. 2023. NOTABLE: Transferable backdoor attacks against prompt-based NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15551–15565, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.867

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems*.

Yisroel Mirsky, Asaf Shabtai, Lior Rokach, Bracha Shapira, and Yuval Elovici. 2016. Sherlock vs moriarty: A smartphone dataset for cybersecurity research. In *Proceedings of the 2016 ACM workshop on Artificial intelligence and security*, pages 1–12. https://doi.org/10.1145/2996758.2996764

R. Molander and Sanyin Siang. 1998. The legitimization of strategic information warfare: Ethical considerations. *AAAS Professional*

Ethics Report, 11(4). https://doi.org/10.7249/MR964

J. H. Moor. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21. https://doi.org/10.1109/MIS.2006.80

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.765

Hellina Hailu Nigatu and Inioluwa Deborah Raji. 2024. ‘‘i searched for a religious song in amharic and got sexual content instead’’: Investigating online harm in low-resourced languages on youtube. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, pages 141–160. ACM. https://doi.org/10.1145/3630106.3658546

Rahil Parikh, Christophe Dupuy, and Rahul Gupta. 2022. Canary extraction in natural language understanding models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 552–560, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.61

Dorian Peters, Karina Vold, Diana Robinson, and Rafael A. Calvo. 2020. Responsible AI—Two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, 1(1):34–47. https://doi.org/10.1109/TTS.2020.2974991

Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2024. A principled framework for evaluating on typologically diverse languages.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021a. ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9566, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.752

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021b. Mind the style of text! Adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.374

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021c. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. 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 443–453, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.37

Fanchao Qi, Yuan Yao, Sophia Xu, Zhiyuan Liu, and Maosong Sun. 2021d. Turn the combination lock: Learnable textual backdoor attacks via word substitution. 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 4873–4883, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.377

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Vyas Raina and Mark Gales. 2022. Residue-based natural language adversarial attack detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, pages 3836–3848, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.281

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.aacl-main.62

Rapid7. 2022. 2022 vulnerability intelligence report. Technical report, Rapid7, Boston, MA.

Amana Raquib, Bilal Channa, Talat Zubair, and Junaid Qadir. 2022. Islamic virtue-based ethics for artificial intelligence. *Discover Artificial Intelligence*, 2. https://doi.org/10.1007/s44163-022-00028-2

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1103

Lena Riecke. 2023. Unmasking the term 'dual use' in EU spyware export control. *European Journal of International Law*, 34(3):697–720. https://doi.org/10.1093/ejil/chad039

Gerald Roche. 2017. Introduction: The transformation of tibet's language ecology in the twenty-first century. *International Journal of the Sociology of Language*, 2017:1–35. https://doi.org/10.1515/ijsl-2017-0001

Gerald Roche and Lugyal Bum. 2018. Language revitalization of tibetan 1. In *The Routledge Handbook of Language Revitalization*, pages 417–426. Routledge. https://doi.org/10.4324/9781315561271-53

Sahar Sadrizadeh, Ljiljana Dolamic, and Pascal Frossard. 2024. A classification-guided approach for adversarial attacks against neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1160–1177, St. Julian's, Malta. Association for Computational Linguistics.

Arthur L. Samuel. 1960. Some moral and technical consequences of automation—A refutation. *Science*, 132(3429):741–742. https://doi.org/10.1126/science.132.3429.741, PubMed: 17797013

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Shruti Sannon and Andrea Forte. 2022. Privacy research with marginalized groups: What we know, what's needed, and what's next. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–33. https://doi.org/10.1145/3555556

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2021.naacl-main.295

Rishabh Singla, Shreyas Srinivasa, Narasimha Reddy, Jens Myrup Pedersen, Emmanouil Vasilomanolakis, and Riccardo Bettati. 2023. An analysis of war impact on ukrainian critical infrastructure through network measurements. In *2023 7th Network Traffic Measurement and Analysis Conference (TMA)*, pages 1–10. https://doi.org/10.23919/TMA58422.2023.10199005

Wim A. Smit. 1995. Science, technology, and the military. *Handbook of Science and Technology Studies. Thousand Oaks (Ca.): Sage*. https://doi.org/10.4135/9781412990127.n26

Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. 2021. Universal adversarial attacks with natural triggers for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3724–3733,

Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.291

Abigail Swenor and Jugal Kalita. 2021. Using random perturbations to mitigate adversarial attacks on sentiment analysis models. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 519–528, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

David Thiel. 2023. Investigation finds AI image generation models trained on child abuse. https://tinyurl.com/swvc493a. Accessed: 2024-06-14.

Andree Thieltges, Florian Schmidt, and Simon Hegelich. 2016. The devil's triangle: Ethical considerations on developing bot detection methods. In *2016 AAAI Spring Symposium Series*.

Nicolas Tournadre. 2013. The tibetic languages and their classification. https://doi.org/10.1515/9783110310832.105

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Olga Tsymboi, Danil Malaev, Andrei Petrovskii, and Ivan Oseledets. 2023. Layerwise universal adversarial attack on NLP models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 129–143, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.10

UNESCO. 2021. Recommendation on the ethics of artificial intelligence.

U.S. Department of Justice. 2022. Department of justice announces new policy for charging cases under the computer fraud and abuse act. https://tinyurl.com/y7een9f8. Accessed: 2024-06-15.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162. Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1221

Jun Wang, Qiongkai Xu, Xuanli He, Benjamin Rubinstein, and Trevor Cohn. 2024a. Backdoor attacks on multilingual machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4515–4534, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.254

Jun Wang, Qiongkai Xu, Xuanli He, Benjamin I. P. Rubinstein, and Trevor Cohn. 2024b. Backdoor attack on multilingual machine translation. https://doi.org/10.18653/v1/2024.naacl-long.254

Yibo Wang, Xiangjue Dong, James Caverlee, and Philip S. Yu. 2024c. DA^3: A distribution-aware

adversarial attack against language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1825, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.107

Zhaoyang Wang, Zhiyue Liu, Xiaopeng Zheng, Qinliang Su, and Jiahai Wang. 2023. RMLM: A flexible defense framework for proactively mitigating word-level adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2757–2774, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.155

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

Norbert Wiener. 1960. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, 131(3410):1355–1358. https://doi.org/10.1126/science.131.3410.1355, PubMed: 17841602

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Tasar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press,

Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pámies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Zongru Wu, Zhuosheng Zhang, Pengzhou Cheng, and Gongshen Liu. 2024. Acquiring clean language models from backdoor poisoned datasets by downscaling frequency space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8116–8134, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.441

Shangyu Xie and Yuan Hong. 2021. Reconstruction attack on instance encoding for language understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2038–2044, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.154

Shangyu Xie and Yuan Hong. 2022. Differentially private instance encoding against privacy attacks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 172–180, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-srw.22

Yong Xie, Dakuo Wang, Pin-Yu Chen, Jinjun Xiong, Sijia Liu, and Oluwasanmi Koyejo. 2022. A word is worth a thousand dollars: Adversarial attack on tweets fools stock prediction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–599. Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.43

Jianhan Xu, Linyang Li, Jiping Zhang, Xiaoqing Zheng, Kai-Wei Chang, Cho-Jui Hsieh, and Xuanjing Huang. 2022. Weight perturbation as defense against adversarial word substitutions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7054–7063, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.523

Yue Xu and Wenjie Wang. 2024. LinkPrompt: Natural and universal adversarial attacks on prompt-based language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6473–6486, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.360

Ying Xu, Xu Zhong, Antonio Jimeno Yepes, and Jey Han Lau. 2021. Grey-box adversarial attack and defence for sentiment classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4078–4087, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.321

Jun Yan, Vansh Gupta, and Xiang Ren. 2023. BITE: Textual backdoor attacks with iterative trigger injection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12951–12968, Toronto, Canada. Association for Computational Linguistics.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021a. RAP: Robustness-Aware Perturbations for defending against backdoor attacks on NLP models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8365–8381, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.659

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021b. Rethinking stealthiness of backdoor attack against NLP 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 5543–5557, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.431

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Biao Yi, Sishuo Chen, Yiming Li, Tong Li, Baolei Zhang, and Zheli Liu. 2024. BadActs: A universal backdoor defense in the activation space. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5339–5352, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.317

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2024. Low-resource languages jailbreak gpt-4.

Ki Yoon Yoo and Nojun Kwak. 2022. Backdoor attacks in federated learning by rare embeddings and gradient ensembling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 72–88, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.6

Wencong You, Zayd Hammoudeh, and Daniel Lowd. 2023. Large language models are better adversaries: Exploring generative clean-label backdoor attacks against text classifiers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12499–12527, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.833

Zhen Yu, Zhenhua Chen, and Kun He. 2024. Query-efficient textual adversarial example generation for black-box attacks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 556–569, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.31

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.540

Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. OpenAttack: An open-source textual adversarial attack toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 363–371, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-demo.43

Yi Zeng, Weiyu Sun, Tran Huynh, Dawn Song, Bo Li, and Ruoxi Jia. 2024. BEEAR: Embedding-based adversarial removal of safety backdoors in instruction-tuned language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13189–13215, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.732

Zhiyuan Zeng and Deyi Xiong. 2021. An empirical study on adversarial attack on NMT: Languages and positions matter. 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 454–460, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.58

Zhiyuan Zhang, Lingjuan Lyu, Xingjun Ma, Chenguang Wang, and Xu Sun. 2022a. Fine-mixing: Mitigating backdoors in fine-tuned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 355–372, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.26

Zhiyuan Zhang, Qi Su, and Xu Sun. 2022b. Dim-krum: Backdoor-resistant federated learning for NLP with dimension-wise krum-based aggregation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 339–354, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.25

Zhuo Zhang, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. 2023. FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9963–9977, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.632

Zeliang Zhang, Wei Yao, Susan Liang, and Chenliang Xu. 2024. Random smooth-based

certified defense against text adversarial attack. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1251–1265, St. Julian's, Malta. Association for Computational Linguistics.

Shuai Zhao, Jinming Wen, Anh Luu, Junbo Zhao, and Jie Fu. 2023. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12303–12317, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.757

Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Yuran Wang, Yong Ding, Yibo Zhang, Qi Zhang, and Xuanjing Huang. 2023. TextObfuscator: Making pre-trained language model a privacy protector via obfuscating word representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5459–5473, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.337

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. 2021. Defense against synonym substitution-based adversarial attacks via Dirichlet neighborhood ensemble. 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 5482–5492, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.426

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Reproducibility of Results

We provide our exact annotations for transparency and reproducibility. The papers sampled in this work are listed in Table 1 (adversarial attacks), Table 2 (backdoor attacks), and Table 3 (Data Reconstruction attacks, which include both embedding inversion attacks and instance encoding/embedding encryption), below. Among these 80 papers, we found zero discussion of dual use and zero instances of coordinated vulnerability disclosure (CVD), so we do not list out the annotations for these two categories.

Paper	Contribution	Ethics	Misuse	Open Source	Languages
Ren et al. (2019)	Attack	No	No	Yes	Eng
Wallace et al. (2019)	Attack	No	No	Yes	Eng
Han et al. (2020)	Both	No	No	Yes	Eng
Zang et al. (2020)	Attack	No	No	Yes	Eng
Li et al. (2020)	Attack	No	No	Yes	Eng
Xu et al. (2021)	Both	Yes	Yes	Yes	Eng
Zeng et al. (2021)	Attack	Yes	Yes	Yes	Eng, Zho
Chen et al. (2021)	Attack	Yes	Yes	Yes	Eng
Song et al. (2021)	Attack	Yes	Yes	Yes	Eng
Li et al. (2021a)	Attack	No	No	Yes	Eng
Zhou et al. (2021)	Defense	No	No	No	Eng
Keller et al. (2021)	Defense	No	No	Yes	Eng
Zeng and Xiong (2021)	Attack	No	No	No	Eng, Zho, Jpn
Swenor and Kalita (2021)	Defense	No	No	No	Eng
Bao et al. (2021)	Defense	No	No	Yes	Eng
Raina and Gales (2022)	Defense	Yes	No	Yes	Eng
Choi et al. (2022)	Attack	No	No	No	Eng, Prog.
Lei et al. (2022)	Attack	No	No	Yes	Eng
Xu et al. (2022)	Defense	No	No	No	Eng
Xie et al. (2022)	Attack	Yes	No	Yes	Eng
Fang et al. (2023)	Attack	Yes	No	Yes	Eng
Cao et al. (2023)	Attack	Yes	No	Yes	Bod
Li et al. (2023c)	Defense	No	No	No	Eng
Wang et al. (2023)	Defense	No	No	No	Eng
Tsymboi et al. (2023)	Attack	No	Yes	Yes	Eng
Gao et al. (2024)	Attack	No	No	No	Eng
Sadrizadeh et al. (2024)	Attack	Yes	Yes	Yes	Eng, Fra, Deu
Zhang et al. (2024)	Defense	No	No	No	Eng
Chen et al. (2024a)	Attack	No	No	No	Eng
Wang et al. (2024c)	Attack	Yes	Yes	Yes	Eng
Xu and Wang (2024)	Attack	Yes	Yes	Yes	Eng
Yu et al. (2024)	Attack	No	No	Yes	Eng
Alshahrani et al. (2024)	Attack	No	No	Yes	Ara

Table 1: Papers sampled pertaining to **adversarial attacks**. Under languages, ‘‘Prog.’’ is short for programming language(s).

Paper	Contribution	Ethics	Misuse	Open Source	Languages
Yang et al. (2021b)	Attack	Yes	No	Yes	Eng
Qi et al. (2021c)	Attack	Yes	Yes	Yes	Eng
Qi et al. (2021b)	Attack	Yes	Yes	Yes	Eng
Qi et al. (2021d)	Attack	Yes	Yes	Yes	Eng
Yang et al. (2021a)	Defense	Yes	Yes	Yes	Eng
Qi et al. (2021a)	Defense	Yes	Yes	Yes	Eng
(Li et al., 2021c)	Defense	Yes	Yes	Empty repo	Eng
Li et al. (2021b)	Attack	No	No	No	Eng
Chen et al. (2022b)	Attack	Yes	Yes	Yes	Eng
Yoo and Kwak (2022)	Attack	No	No	No	Eng
Gan et al. (2022)	Attack	Yes	Yes	Yes	Eng
Chen et al. (2022a)	Defense	Yes	No	Yes	Eng
Zhang et al. (2022a)	Defense	No	No	No	Eng
Lyu et al. (2022)	Defense	No	No	Yes	Eng
Jin et al. (2022)	Defense	Yes	Yes	Broken link	Eng
Zhang et al. (2022b)	Defense	Yes	No	No	Eng
Liu et al. (2023)	Defense	No	No	No	Eng
Zhao et al. (2023)	Attack	Yes	Yes	Yes	Eng
Mei et al. (2023)	Attack	Yes	Yes	Yes	Eng
He et al. (2023a)	Defense	No	No	Yes	Eng
You et al. (2023)	Both	No	No	No	Eng
Li et al. (2023d)	Attack	Yes	Yes	Yes	Eng, Prog.
Yan et al. (2023)	Both	Yes	Yes	Yes	Eng
Li et al. (2023b)	Defense	Yes	Yes	Yes	Eng
He et al. (2023b)	Defense	No	No	Yes	Eng
Huang et al. (2024a)	Attack	Yes	Yes	Yes	Eng
Li et al. (2024)	Attack	Yes	Yes	Yes	Eng
Du et al. (2024)	Attack	Yes	Yes	No	Eng
Graf et al. (2024)	Defense	Yes	Yes	Empty repo	Eng
Zeng et al. (2024)	Defense	Yes	Yes	Broken link	Eng, Prog.
Yi et al. (2024)	Defense	Yes	No	Yes	Eng
Wu et al. (2024)	Defense	Yes	No	Yes	Eng
Wang et al. (2024a)	Attack	No	No	No	Eng, Jav, Ind, Msa, Tgl, Tam

Table 2: Papers sampled pertaining to **backdoor attacks**. Under languages, ‘‘Prog.’’ is short for programming languages.

Paper	Contribution	Ethics	Misuse	Open Source	Languages
Huang et al. (2020)	Defense	No	No	Yes	Eng
Xie and Hong (2021)	Attack	Yes	Yes	No	Eng
Xie and Hong (2022)	Defense	No	No	No	Eng
Hayet et al. (2022)	Attack	No	No	Yes	Eng
Parikh et al. (2022)	Both	Yes	Yes	No	Eng
Kim et al. (2022)	Defense	No	No	No	Eng
Morris et al. (2023)	Attack	No	No	Yes	Eng
Li et al. (2023a)	Attack	Yes	No	Yes	Eng
Zhou et al. (2023)	Defense	No	No	Yes	Eng
Zhang et al. (2023)	Defense	No	No	Yes	Eng
Chen et al. (2024b)	Both	Yes	Yes	Yes	Eng, Fra, Deu, Spa
Huang et al. (2024b)	Attack	No	No	Empty repo	Eng
Elmahdy and Salem (2024)	Attack	No	No	No	Eng
Lin et al. (2024)	Attack	No	No	No	Eng, Fra, Deu, Spa, Ell, Bul, Rus, Tur, Ara, Vie, Tha, Zho, Hin, Swa, Urd

Table 3: Papers sampled pertaining to **data reconstruction attacks**, which includes both **embedding inversion** and **embedding encryption** (i.e., instance encoding).

B Complementary Results

In this Appendix, we provide more general information about the sampled papers, for further transparency. First, Figure 8 shows the relative age of publications, and Figure 9 shows their distribution across venues within the ACL* community. Figure 10 presents the continent associated with every unique affiliation in the author list, so that we may provide rough demographic information about our sample. Finally, Figure 11 show the most common evaluation datasets across our sample of works in NLPSec.

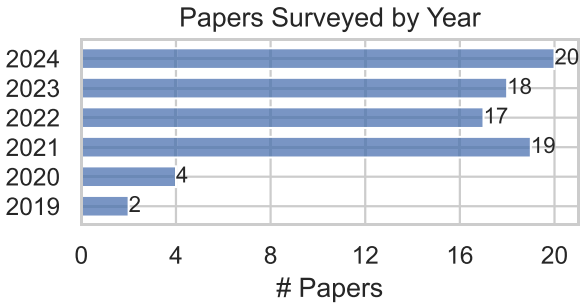

Figure 8: Distribution of sampled papers across their year of publication.

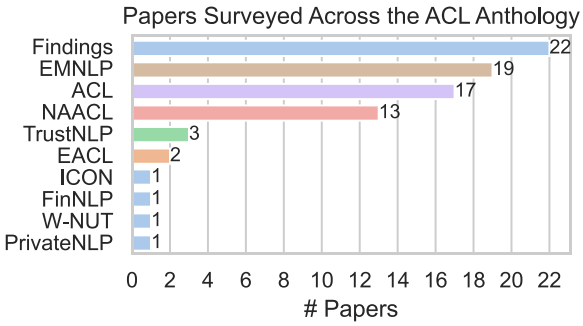

Figure 9: Distribution of sampled papers published in different venues across the ACL* community. For the Findings papers specifically, 11 are at ACL, 8 at EMNLP, 2 at NAACL, and 1 at EACL.

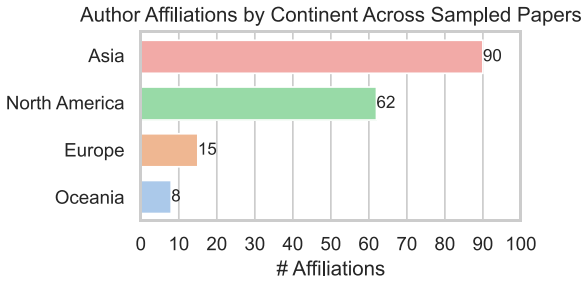

Figure 10: We show the distribution of author affiliations across continents, rather than countries, for easy comparison against Kaffee et al. (2023). Most research from our sample of NLPSec works hails from institutions residing in Asia. Conversely, governance documents come overwhelmingly from the EU and US (Jobin et al., 2019), and discussions on dual-use and ethics in NLP are also Western-centric (n = 3 from Asia and n = 1 from Africa in a survey of 48 people by Kaffee et al., 2023). This highlights the need for a wider dialogue across the field. When cross-cultural values conflict and best practices become unclear, practitioners should consider the UN Declaration of Human Rights.

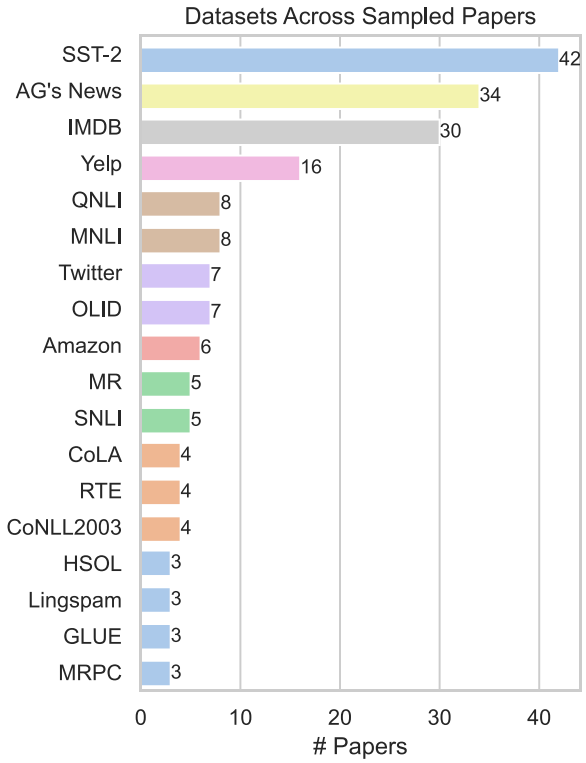

Figure 11: Distribution of datasets used to validate experiments across the sampled works in NLPSec. For the purposes of visualization, we do not display the long tail of datasets which were used by only one paper.