Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset

Peter Henderson[*], **Mark S. Krass**[*], **Lucia Zheng, Neel Guha**
Christopher D. Manning, Dan Jurafsky, Daniel E. Ho
Stanford University

Abstract

One concern with the rise of large language models lies with their potential for significant harm, particularly from pretraining on biased, obscene, copyrighted, and private information. Emerging ethical approaches have attempted to filter pretraining material, but such approaches have been ad hoc and failed to take context into account. We offer an approach to filtering grounded in law, which has directly addressed the tradeoffs in filtering material. First, we gather and make available the Pile of Law, a ∼256GB (and growing) dataset of open-source English-language legal and administrative data, covering court opinions, contracts, administrative rules, and legislative records. Pretraining on the Pile of Law may help with legal tasks that have the promise to improve access to justice. Second, we distill the legal norms that governments have developed to constrain the inclusion of toxic or private content into actionable lessons for researchers and discuss how our dataset reflects these norms. Third, we show how the Pile of Law offers researchers the opportunity to learn such filtering rules directly from the data, providing an exciting new research direction in model-based processing.

Warning: this paper contains quotations that may be offensive or upsetting.

1 Introduction

The presence of private and toxic content in the most popular corpora for pretraining large language models is a well-known problem [11, 58]. But what to do about it is largely a matter of researcher discretion. Some teams implement extensive processes for filtering content deemed toxic or private; others train on data in virtually unmodified form. Resolving all of the difficulties and nuances of content filtering can be challenging, potentially explaining why content filtering has been so uneven.

It is practically difficult to perform reliable and transparent filtering at scale. That is partially because undesirable content is deeply contextual. For example, whether the inclusion of a racial epithet in a dataset is toxic may depend on factors such as the identity of the speaker and the expectations of the readers [43, 114]. Likewise, the existence of privacy violations may depend in part on the extent to which a speaker expected a fact to be widely shared at the time it was expressed [20, 6-7]. And privacy expectations may vary widely across countries [10].

Any filtering process involves complex trade-offs. Filtering for toxicity may have unexpected effects on representation in datasets or the bias of downstream outputs [40, 48, 6]. And filtering too widely for privacy may harm important downstream applications, as when the Census Bureau's adoption of differential privacy led to errors in redistricting U.S. Congressional districts [102].

Yet researchers are not the first to balance the merits of open-source transparency with potential harms: legal and administrative actors have expended significant resources and process in developing

[*]Equal contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

https://doi.org/10.52202/068431-2119

standards to strike this exact balance. In this work, we suggest that researchers can look to these long-developed (and debated) standards to help ground content filtering mechanisms for large language model training.

This paper makes three contributions. First, we curate and open-source a ∼256GB (and growing) dataset of legal and administrative data, which we call Pile of Law, which can be used for assessing norms on data sanitization across legal and administrative settings. This dataset can be an exploratory tool for evaluating different mechanisms for "doing the data work" [104]. And we note that pretraining on the Pile of Law may help with challenging legal tasks that have the potential to improve access to justice [16]. Second, we catalog how government has, though extensive legislation, regulation, and litigation, developed standards for handling the trade-offs between privacy and offensive content on the one hand and transparency, access, and completeness on the other. We suggest actionable insights for researchers based on these legal and administrative norms.[2] Third, we demonstrate how implicit sanitization rules can be learned from the Pile of Law, providing a path forward for researchers to develop more nuanced filtering mechanisms. We also demonstrate shortcomings in alignment for current sanitization techniques, providing exciting new directions for research.

2 Pile of Law

We curate a ∼256GB (and growing) dataset of legal and administrative text.[3] The utility of this data is twofold: (1) to aggregate legal and administrative data sources that demonstrate different norms and legal standards for data filtering; (2) to collect a dataset that can be used in the future for pretraining legal-domain language models, a key direction in access-to-justice initiatives [16]. A number of prior works have pretrained smaller models on smaller subsets of legal data, including private data that is subject to restrictive licenses [30, 129]. None of these have conducted an analysis of the legal data itself—and none have curated an open-source, legal-focused pre-training dataset at this scale.

Through extensive efforts, we compile data from 35 data sources, including legal analyses, court opinions and filings, government agency publications, contracts, statutes, regulations, casebooks, and more. Others have aggregated smaller subsets of legal data, such the EuroParl datasets which gather European Parliamentary debates [64, 57]. We have included some of these as subsets of Pile of Law when relevant and plan to continue adding material to the Pile of Law over time, further increasing its utility to the community.

We characterize the dataset in detail in Appendix E. All of the content is already entirely public and mostly available under permissive licenses, but has not previously been compiled at scale for research purposes.[4] Each of these data sources carries with it an implicit filtering mechanism formed under relevant legal standards of privacy and toxicity, which we discuss throughout subsequent sections and in the Appendix. While the underlying data in Pile of Law is already public record and has implicit filters, we recognize that it may contain sensitive material that has escaped administrative scrutiny. We discuss the ethics of our work and our proposed mechanisms for content removal in Appendix A.

This dataset has obvious utility for pretraining legal-domain foundation models, particularly since, unlike other pretraining data, all material is under open licenses. Though not central to our work, we demonstrate this potential by training an initial BERT-large equivalent model on Pile of Law, yielding comparable results to highly context specific (but smaller) models (see Appendix F for full results). Recent research has shown in legal contexts that pretraining smaller models on highly in-domain data may be better than large models on big data [129, 29]. But in theory, there should be generalizable knowledge and skills that can be learned by training across more diverse sources of data. A well-crafted pretraining procedure that instills analogical reasoning abilities, for example, should transfer across domains. Our dataset is large and diverse enough (covering distinct areas of law like criminal law, contracts, and administrative law) to test this hypothesis in the legal domain, where our initial models can form a baseline.

[2] Note: while we discuss a number of privacy and toxicity standards, due to the expertise of the authors and the availability of data, this work focuses on the U.S. legal system. We address this and other limitations in Appendix B.

[3] https://huggingface.co/datasets/pile-of-law/pile-of-law.

[4] See Appendix G for a discussion of copyright and licensing in the dataset.

https://doi.org/10.52202/068431-2119

Table 1: Filters Applied in Major Pre-Training Papers

	PSI	Deduplication	Toxic Content	Quality
CCNet [125]	No	MinHash (pages)	No	No
C4 [98]	No	Unknown (3-sentence spans)	Word list	Minimum word counts, presence of curly brackets, 'lorem ipsum', etc.
GPT-3 [21]	No	MinHash (pages)	No	Train classifier to distinguish CC from curated high-quality examples
Gopher [97]	No	MinHash (pages)	Google SafeSearch	Min./max. word counts, word-to-symbol ratio, share ellipses, excessive repetition; require stop words
The Pile [44]	No	MinHash (pages)	Ad-hoc source deletion	Train classifier to distinguish CC from curated high-quality examples

3 What Can the Law Teach Us About Content Filtering?

When releasing internal documents concerning individuals, courts and governments have long struggled to balance transparency against the inclusion of private or offensive content. Model creators now face a similar struggle: what content to filter before pretraining a large language model on the data. In this section we survey how governments and courts have handled such content filtering and briefly discuss how Pile of Law implicitly encodes these privacy and toxicity rules. Based on these rules, we provide actionable lessons for researchers training large language models across fields, so that they can adapt similar rules as minimum standards for dataset sanitization. To be clear, we do not take the position that legal rules are optimal nor monolithic. But in many cases they result from a deliberative process that includes judges, legislators, and policymakers in contexts open to public scrutiny, so we think that the machine learning community can at minimum learn from these laws, rules, and norms to improve current ad hoc practice. In short, there is no need to reinvent law.

3.1 Privacy

Despite the growing focus on privacy in NLP [20], Table 1 shows that many major pre-training papers do not explicitly filter potentially sensitive information (PSI).[5] For example, [44] excludes sources due to concerns over explicit or racist content, but does not assess the prevalence of PSI, despite including web-based sources (e.g. OpenWebText) in which users may have an expectation of anonymity. Instead, pre-training papers have focused their attention on alternatives to filtering, like deduplication [62], federated learning [108, 50], differential privacy [80, 42], and other approaches [32, 74, 65, 127, 82]. But a number of recent papers have demonstrated that large generative models output memorized content [26, 111, 32, 25, 69] even with deduplication [27]. Given that many models are trained without privacy mechanisms, filtering is critical to protecting individuals, which is perhaps why research involving health data still emphasizes that approach [90, 2]. But choosing what to filter is challenging; below, we discuss how governments and courts make such decisions.

How have governments balanced privacy against competing values? First, we examine how several jurisdictions handle privacy filtering. Table 2 provides a brief summary.[6]

Baseline Redactions. Across the jurisdictions we examine, there is a baseline level of filtering. Virtually every jurisdiction in Table 2 protects the identities of minors. At minimum, juveniles must be protected by pseudonyms in public judgments, and outside of some U.S. states, juvenile criminal records are not public. No jurisdiction normally permits the publication of financial account numbers,

[5] We define PSI to mean information that could violate a person's privacy interests. This could include personally identifiable information, including a person's name, date of birth, or identification number. Under this definition, a document can contain some PSI (e.g. a name or the facts of a case) while excluding other PSI (e.g. date of birth). But some information that is personally identifiable is not PSI; for example, the name and office contact information of an attorney filing a court brief is identifying but not sensitive.

[6] See Appendix I for a complete version of the table, including citations.

https://doi.org/10.52202/068431-2119

Table 2: Availability of Identifying Information Across Administrative Settings

Jurisdiction	Civil Cases	Criminal Cases	Juvenile Data
U.S. Federal Courts	All case details public unless sealed, except DOBs, ID/account #s.	Def. names public; DOBs, ID/account #s, addresses redacted.	Criminal records confidential. Names redacted from civil cases.
U.S. Admin. Agencies	Most PII omitted from public records.	-	No statute; more protection in practice.
German Courts	Judgments omit all identifying information.	Confidential 3-5 years after sentence completed.	No public access to criminal records.
Chinese Courts	Names/case details public except in certain classes of cases.	Names/case details are public as of 2016.	Juvenile criminal records are categorically exempt from disclosure.
Canadian Courts	Presumption of openness, except specific details and rare sealed cases.	Public; may be sealed after a period of good behavior.	Youth criminal records are always confidential.

dates of birth, or identity numbers like social security numbers.[7] All of these are bright line rules directly applicable to text corpora.

Value-system contexts. There are also significant points of disagreement corresponding to the role of privacy in different value systems. U.S., Chinese and Canadian courts denote the names of litigants in ordinary civil cases, prioritizing public access and transparency; German courts do not. Likewise, U.S. federal courts virtually never remove criminal cases from the public record [109, p. 1233], a rule also emerging in China [75]. Canada allows most criminal records to be expunged after a period of good behavior. And in Germany, virtually all criminal records are automatically sealed after a set time, and courts have even imposed fines for publicizing a person's criminal history after expungement [22].

Contextualized privacy. Digging further into these rules highlights how court privacy rules account for context. In the U.S. and Canada, the public disclosure of litigants' potentially sensitive information (PSI) can be avoided by persuading a court that extenuating circumstances apply [124, 115]. To name one example, courts generally permit pseudonyms when parties allege that they have suffered a sexual assault [124, p. 57]. The chance to *seal* a case, or to file pseudonymously, suggests that even the most open judicial regime allows for censoring in exceptional cases—although the sealing and pseudonymous filing standards suffer from inconsistency and misuse [113].

Likewise, administrative agencies often employ context-aware heuristics when deciding whether to include PSI in public decisions. Although administrative courts are not generally bound by stringent privacy rules like HIPAA [110, 36], the Department of Justice exempts immigration applications from public scrutiny due to privacy concerns [87]; the same is true of Social Security Disability applications. Cases involving veterans' benefits are released pseudonymously.

Public availability is not a limit. In many cases, the rules for sanitizing PSI and sealing cases do not depend on whether information is already public. For example, the ban on publicly filing documents revealing dates of birth in U.S. federal courts does not depend on whether a litigant's birth date is otherwise public [122]. In cases where a court does take into account the public availability of information (e.g., sealing standards [115, 121]), contextual countervailing factors can justify keeping a case sealed.

Implications for Pile of Law. All of the above privacy norms mean that each subset of Pile of Law is already filtered for privacy based on legal norms in that jurisdiction. Further filtering could seek to align the whole dataset with the norms in one of the subsets prior to pretraining. Appendix E summarizes the filtering norms present in each subset of the data.

[7] The United States' Federal Rule of Civil Procedure 5.2 lays out exceptions when these facts are contained in judicial records properly before a federal court and for civil asset forfeiture cases.

Lessons for researchers. First, the law provides a number of useful heuristics that researchers could deploy to sanitize data. Detecting and redacting juvenile names, dates of birth, and account and identity numbers is virtually always appropriate across countries. Legal protections for already-public information show why sanitization may be necessary even for text collected from public-facing web pages. Second, the U.S. system appears to lean most heavily toward transparency. We suggest that researchers can use the U.S. court rules as a floor. Such privacy filtering rules would already go beyond much of current modeling practice. Third, in addition to consensus heuristics, researchers should make contextualized decisions about privacy harms. While this may seem difficult, Section 4 demonstrates how to leverage Pile of Law to learn contextualized standards to mimic legal privacy redaction mechanisms; alternatively, allowing individuals whose information appears in the training corpus to request removal may serve as another stopgap. Last, the U.S. legal rules do not extend as far as some researchers suggest is necessary. For example, [5] suggests that *all* names must be redacted to preserve privacy. This would reflect greater privacy protection than is typically afforded by U.S. law, which prioritizes public openness and transparency about court proceedings, but would be in line with German rules. These pose important value tradeoffs, and we suggest that researchers look as a starting point to the jurisdiction that aligns with such value tradeoffs for filtering other potentially sensitive content.

3.2 Toxicity

How is toxic speech defined in research? The category of 'toxic speech' is defined in multiple ways [47, 123, 4]. Some papers define toxicity as "*disrespectful* comments, including . . . identity *attacks*, profanity and *threats*," thus emphasizing the idea of intentional insult [39, 126, 128]. A broader definition would incorporate *implicit* toxicity, as when a speaker "subtly" or "unconsciously expresses a prejudiced attitude" [18, 70]. [18] cites the example of the question "But where are you from, originally?" Others would take a still broader view, suggesting that any *profanity* is toxic (in addition to hate speech and derogatory content) irrespective of speaker intent [89]. One implication of these divergent choices concerns *mentions* of toxic language, where a speaker refers to something said by another [114]. For example, if a judge writes that "Plaintiff claims that her supervisor called her '___"' (where ___ is a profane epithet), an intent-based standard typically would not deem the use of ___ 'toxic,' while an approach targeting profanity typically would.

How have governments regulated toxic content? Scholars have documented the role of the law in institutional racism and other forms of oppression, and legal materials from prior eras use words that would by modern standards be considered epithets [31, 13]. Today, the legal profession in most Anglophone countries strongly polices overt discriminatory epithets [38]. Overtly biased speech is prohibited for judges and lawyers in the U.S., Canada, and the U.K. by professional rules [3, 24, 119, 68]; similar norms have been put forward by the U.N. [120]. Judges and lawyers in all countries are routinely sanctioned when they use racist epithets, and most incidents occur verbally or off the bench [38, 60, 118].

Unlike overt, indecorous epithets, legal norms permit the use of speech affected by implicit bias; the incidence of such speech is well-documented [94, 100, 85, 12]. Indeed, it is sometimes encoded in the laws judges and administrative agencies enforce [23]. Furthermore, some lawyers may see themselves as professionally *obligated* to deploy stereotypes when doing so may assist their clients (e.g. immigration [88], defendants in sexual assault cases [34]).

Implications for Pile of Law. The adversarial legal system in many Anglophone countries creates incentives for lawyers to complain about overt racism in written materials, which would violate unambiguous professional rules. Thus, the appearance of epithets in our data is more likely to be confined to quotations, mentions, or to historical legal materials. However, text in our corpus may be toxic according to other definitions; for example, we are unable to quantify the prevalence of implicit biases or offensive stereotyping. Explicit racial, sexual, or offensive terms do appear in modern legal text, but most often in the form of a quotation than direct use. For instance, many cases revolve around evidence documenting racial or gender discrimination, and judges commonly spell out profane or explicit words from the evidentiary record [63, 45]. Finally, elected officials in our legislative transcripts are not bound by the same professional norms as attorneys. Additionally, an interesting future examination may note differences between civil law and common law systems,

https://doi.org/10.52202/068431-2119

examining rates of offensive content between the different legal systems and norms. We provide a per-subset examination of filtering norms in Appendix E.

Lessons for researchers. First, as is true of privacy, the toxicity norms prevalent in many legal systems offer a lower-bound for researchers. Researchers seeking to mimic the standards that apply in courts should sanitize intentional uses of derogatory terms from pretraining data. That said, current filters are not precise enough to handle this standard. Under the rules applicable to lawyers, filters based on simple word lists would be over-inclusive because they would capture *references* to offensive language that may be non-toxic in context. Second, the rules applied in courts suggest that generative models should portray toxic behavior explicitly in some contexts, either to serve the values of 'accuracy and precision' or to persuade readers [63, p. 7]; but as [43] argues, this view is contested.

Third, in language model pretraining, there may be reason to exceed minimum judicial standards depending on the length of content needed to contextualize references to offensive speech. Accessible language models like Roberta [78] have a maximum context window of 512 tokens. If a reference to offensive content spans the majority of these tokens, the model will simply uptake the offensive content as if it were being trained for *direct use.* As model contexts grow, it may become more reasonable for researchers to adopt judicial norms.

4 What Can We Learn from Legal Text?

As Section 3 shows, even jurisdictions that impose a strong presumption of transparency on legal documents often allow for contextual decisions that weigh this presumption against the potential harms caused by the inclusion of PSI on the public record. Reducing these rules to tools that can be deployed for filtering may be challenging. But Pile of Law encodes these contextual decisions already, providing a rich opportunity to learn context-aware norms directly. This section demonstrates the promise of Pile of Law for operationalizing legal norms. While not comprehensive, the experiments below demonstrate a path forward for replicating the content-filtering mechanisms of courts and governments by leveraging variation in Pile of Law. In particular, we show that: (1) Pile of Law reflects variation in privacy norms that can be leveraged to learn contextual privacy rules, such as when to redact names in potentially harmful situations; (2) Pile of Law reflects variation in toxicity norms over time and across contexts, toxicity filters fall short of handling these nuances, and researchers can learn much from building toxicity filters that can handle nuances in Pile of Law's text.

4.1 Learning Contextual Privacy Rules

Case Study 1: Pseudonymity in Immigration Court. The Board of Immigration Appeals (BIA) evaluates petitions appealing immigration decisions and sometimes publishes precedential decisions that affect future cases. Some cases include applicants' full names, while others replace them with pseudonymous initials. We demonstrate how subsets of the data can be used to learn the value judgements made in making this pseudonymity decision. We split cases into paragraphs and mask terms used to refer to the applicant. We train a distill-BERT base model [105] to predict whether the paragraph should use pseudonymity or not. This model achieves $\sim$80% F1 on the validation set. We then examine what types of content are more likely to trigger a pseudonymity recommendation by conducting a perturbation analysis. We use the Bias in Bios dataset [35], censored for names and pronouns. We prepend an additional sentence to each biography that indicates whether the person: (1) is seeking asylum or is a refugee; (2) experienced torture; (3) committed a non-violent criminal offense; or (4) committed a violent criminal offense. Figure 1 shows that asylum and torture sentences were more likely to trigger pseudonymity while criminal offenses were less likely. This aligns with federal regulations that prevent disclosure of information related to asylum or the Convention Against Torture (8 CFR § 208.6(a)). By contrast, federal regulations allow information disclosure when a criminal proceeding is involved (8 CFR § 208.6(d)(1)(ii)), though no regulation addresses criminal history.

Next, we fit a causal lexicon using the deep residualization method (and associated library) from Pryzant et al. [95]. We control for the year that a case was published since we found that some aspects of privacy standards have shifted year-to-year, which provides a unique opportunity to learn evolving standards of privacy. We select the top 100 most indicative terms for pseudonymity and remove those

https://doi.org/10.52202/068431-2119

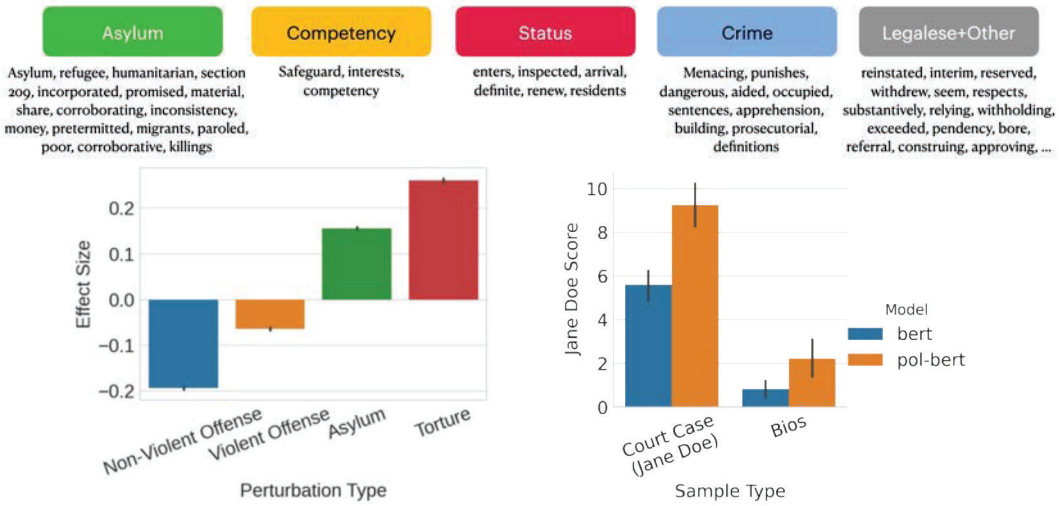

Figure 1: (Top) A causal lexicon learned for the EOIR privacy task, manually sorted by topic with contextual information. Extended version and information in Appendix H.1. (Bottom Left) A distill-bert model is more likely to predict pseudonymity for bios with an asylum or torture perturbation (effect size is difference in pseudonymity likelihood from normal bio and bio with added perturbation). (Bottom Right) Jane Doe Score is the difference in MLM score between a version of the sentence using Jane Doe and a random name. The sample sources are paragraphs using pseudonyms and Bios [35] (no pseudonyms).

where the term only showed up in one case. Then we manually examine contexts and cluster terms into categories. We found that terms most likely to be associated with pseudonymity could be largely clustered into: asylum, mental competency (a legal term used to refer to one's ability to stand trial), immigration status, and indications of a criminal proceeding. We also find that many terms associated with general legal language were included, suggested some remaining confounding and the need for more research into text-based causal attribution. These causal lexicons are seen in Figure 1.

Case Study 2: Pseudonyms in Civil Litigation. Next, we look to a "zero-shot" version of the experiment above in a broader setting. As noted in Section 3, litigants in U.S. courts can ask to use pseudonyms like "Jane Doe" in court documents, for example in harassment suits. To assess these requests, courts consider contextual factors like "sensitive and highly personal" subject matter, minors, or other extenuating circumstances [124]. We collect $\sim$ 500 paragraphs where a pseudonym ("Jane Doe" or "Jane Roe") is used from the validation part of the Court Listener Opinions data. For each sentence, we create 100 alternative sentences that replace "Jane Doe" with a name sampled using 1990 Census probabilities (using NAMES). We then compare whether each model is more likely to guess "Jane Doe" using MLM Score [103]. We repeat this process on the Bios dataset [35]. Figure 1 shows that a model trained on Pile of Law (pol-bert) ranks Jane Doe $\sim$ 3 points higher than a standard bert-large-uncased on true pseudonym cases. This suggests that models pre-trained on Pile of Law are more likely to encode appropriate pseudonymity norms. To be sure, pol-bert is slightly more biased for Jane Doe use overall, as is to be expected, but its performance gains persist even after accounting for this bias.

Case Study 3: Privacy Standards in Medical Cases. We examine *inter-source variation* between the Board of Veterans Appeals (BVA) and the Department of Labor's Employee's Compensation Appeals Board (DOL). Leading tools for data sanitization remove personal health information as defined by HIPAA [33], including dates or the name of a physician [90]. We ran [90] on all decisions by the BVA and DOL since both adjudicate the extent of applicants' disabilities, though they are not bound by HIPAA [36]. Showing the difficulty of applying sanitization tools out of domain, virtually *all* decisions included information flagged as HIPAA-protected: 99% included dates; 96% of BVA and 100% of DOL decisions included medical facility names. But the two agencies also differed. About 26% of DOL cases but just 0.36% of BVA cases included a physician name. Physician fraud is more common in worker's compensation programs like DOL's [91], but the BVA relies on the

https://doi.org/10.52202/068431-2119

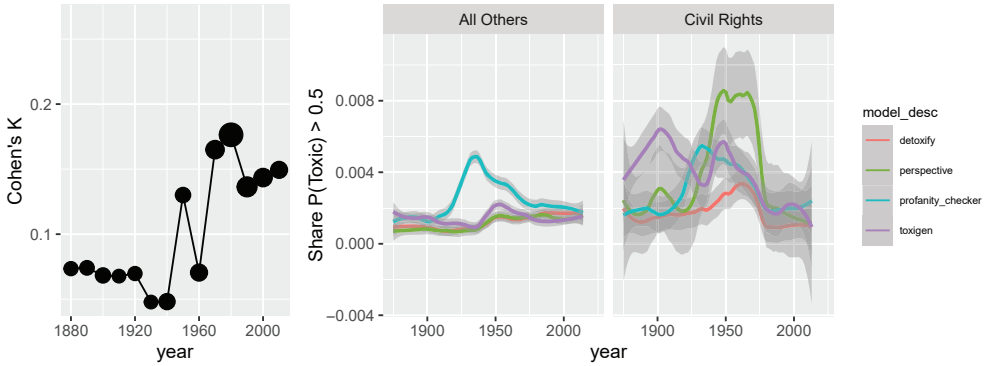

Figure 2: Inter-Model Agreement and Toxicity Over Time. Left: Cohen's κ, by 10-year bin, calculated for [130] and [59], with dot size proportional to number of examples. Right: Share of sentences assigned >50% probability of being toxic, by model, time, and topic classification [112].

testimony of VA physicians. The transparency in DOL opinions reflects the higher public interest in physician accountability.

Lessons for researchers. These experiments show that the Pile of Law encodes signals about privacy standards that can be learned to produce more nuanced recommendations about filtering. For example, researchers may consider whether to mimic the EOIR standard to remove names in proceedings related to minors, asylum or safety concerns. Or they may wish to learn and apply the more contextual standard that is used in general U.S. litigation, where a complex set of factors is used to justify the exclusion of names from case texts. Such contextualized filters may help ensure that generative models strike the right balance between accuracy and privacy protection, for example by accurately distinguishing benign releases of names and contact information (e.g., in response to queries about government officials) from harmful ones (sensitive circumstances where harm is plausible).

4.2 Calibration and Value-Alignment in Toxicity Filtering

We also identify three main insights (and challenges) from using toxicity filters on Pile of Law, setting the ground for new research using the dataset: (1) toxicity filters often disagree, creating potential issues for automated filtering; (2) toxicity filters may be value-misaligned when it comes to content that is flagged in Pile of Law; (3) toxicity scores vary highly with the length of the content, making it unclear how to handle long-document filtering.

Case Study: Supreme Court Decisions. Leveraging Pile of Law, we show that there are profound nuances to filtering toxic content. First, toxicity filters encode value judgements and divergent definitions of toxicity. Figure 2 shows Cohen's κ between profanity-checker and Perspective over time for sentences in Supreme Court cases (Fig. 6 shows the same for all filters). At the sentence level, the tools' agreement rates are very low, but rise over time, indicating the challenge of handling out-of-domain data far away in time. A vivid example of this challenge is provided in Table 3: *Dred Scott* is the most notoriously racist decision in U.S. history [46], but perhaps due to the archaic language of its holding, *none* of the models is sure that it is toxic.

But civil rights cases illustrate why the disagreement is about conceptual differences, not just domain drift. Figure 2 shows that the period between 1950 and 1970 is associated with a large spike in the share of sentences deemed toxic in U.S. civil rights cases. This period was associated with the end of *de jure* segregation in the United States [117]. Many cases likely *quoted* or *mentioned* racist laws before striking them down. For instance, in Table 3, *Loving* describes a law banning interracial marriage in order to deem it illegal. Quoting this language qualifies as toxic under some but not all definitions, and as Figure 2 shows, that view is encoded in some but not all filters. Accordingly, the filters disagree as to whether *Loving*'s quote is clearly toxic. Document-level filtering could thus easily delete core civil rights cases like *Hunter* and *Loving*—while leaving in *Dred Scott*.

Finally, we note that the context window used to filter out sentences appears to dramatically influence ratings. Perspective segments data into sentences and then labels each sentence, which is the

Table 3: Toxicity Ratings of Quotes From the U.S. Supreme Court, Showing Rating Disagreement

Case	Quote	(1)	(2)	(3)	(4)
Hunter v. Erickson (1969)	"The majority needs no protection against discrimination."	0.02	0.05	0.00	0.81
Loving v. Virginia (1967)	"[I]f any white person intermarry with a colored person . . . he shall be guilty of a felony and shall be punished by confinement in the penitentiary . . .."	0.52	0.54	0.60	0.94
Dred Scott v. Sandford (1857)	"A free negro of the African race whose ancestors were brought to this country and sold as slaves is not a citizen within the meaning of the Constitution."	0.29	0.50	0.26	0.54

Note: Model **(1)** is profanity-check [130]; **(2)** is Perspective [59]; **(3)** is Detoxify [49]; and **(4)** is Toxigen [51] .

approach we take above. We find that by using longer span, we can *systematically decrease* the perceived toxicity of a span, even if it is obviously toxic under any definition. We take the top 5k sentences labeled as toxic by Toxigen. We then take 2 sentences before and after the toxic sentence (clamped to the boundaries of the document). We find that the toxicity score drops between **55-57%** (absolute, 95% CI) just by adding this context. While some of this change might be due to correct re-classification of mentions, we provide qualitative examples in which this is clearly untrue in Appendix Table 8.

Lessons for researchers. The experiments above demonstrate that, while toxicity filtering is important to align with the courts' modern lower bounds banning uses of epithets, it is not clear that existing filters are not consistent and filter out content aligned with different values. Moreover, they can arbitrarily label content as non-toxic in long-document or out-of-distribution settings, which may affect filtering mechanisms. More work is needed to create robust, value-aligned toxicity filters for pretraining and it is unclear if off-the-shelf mechanisms strike the right balance. As we have shown, the Pile of Law provides unique opportunities to develop such methods.

5 Conclusion

In this work we have examined how the law and legal data can inform data filtering practices that are of great importance to responsible large language model training. We provide an extensive legal dataset (the Pile of Law) and illustrate a number of exciting new research directions for future work.

Acknowledgements

We thank SambaNova Systems for generously providing compute resources via the SambaNova Systems Dataflow-as-a-Service™ platform and the Stanford Institute for Human-Centered Artificial Intelligence for computing support. We also thank Jieru Hu for helpful discussions and Krithika Iyer for technical assistance. PH is supported by an Open Philanthropy Project AI Fellowship.

References

[1] Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the Limits of Large Scale Pre-training. In *International Conference on Learning Representations*, 2021.

[2] Allison Adams, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. AnonyMate: A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, 2019.

[3] American Bar Association. Model Code of Judicial Conduct, 2007. https://www.americanbar.org/groups/professional_responsibility/publications/model_code_of_judicial_conduct.

https://doi.org/10.52202/068431-2119

[4] Noman Ashraf, Arkaitz Zubiaga, and Alexander Gelbukh. Abusive language detection in youtube comments leveraging replies as conversational context. *PeerJ Computer Science*, 7: e742, 2021.

[5] Tuomas Aura, Thomas A Kuhn, and Michael Roe. Scanning electronic documents for personally identifiable information. In *Proceedings of the 5th ACM workshop on Privacy in Electronic Society*, pages 41–50, 2006.

[6] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 15479–15488, 2019.

[7] Azeem Bande-Ali and Walker Boyle. U.S. Supreme Court annotated transcripts. Github, 2019. https://github.com/walkerdb/supreme_court_transcripts.

[8] Alexander Baturo, Niheer Dasandi, and Slava J Mikhaylov. Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics*, 4(2), 2017.

[9] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit Dataset. In *14th International Conference on Web and Social Media*, pages 830–839, 2020.

[10] Steven Bellman, Eric J. Johnson, Stephen J. Kobrin, and Gerald L. Lohse. International Differences in Information Privacy Concerns: A Global Survey of Consumers. *The Information Society*, 20(5):313–324, 2004.

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

[12] Janine Benedet. Judicial Misconduct in the Sexual Assault Trial. *U.B.C. L. Rev.*, 52:1, 2019.

[13] Mary Frances Berry. *The Pig Farmer's Daughter and Other Tales of American Justice: Episodes of racism and sexism in the courts from 1865 to the present*. Vintage, 2011.

[14] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Tax Law NLP Resources, 2020. https://doi.org/10.7281/T1/N1X6I4.

[15] Michael J Bommarito II, Daniel Martin Katz, and Eric M Detterman. LexNLP: Natural language processing and information extraction for legal and regulatory texts. In *Research Handbook on Big Data Law*. Edward Elgar Publishing, 2021.

[16] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. *arXiv e-prints*, art. arXiv:2108.07258, August 2021.

[17] Łukasz Borchmann, Dawid Wisniewski, Andrzej Gretkowski, Izabela Kosmala, Dawid Jurkiewicz, Łukasz Szałkiewicz, Gabriela Pałka, Karol Kaczmarek, Agnieszka Kaliska, and Filip Graliński. Contract Discovery: Dataset and a Few-Shot Semantic Retrieval Challenge with Competitive Baselines. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4254–4268. Association for Computational Linguistics, November 2020.

[18] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China, November 2019. Association for Computational Linguistics.

[19] British Columbia Law Society. Code of Professional Conduct for British Columbia, 2021. https://www.lawsociety.bc.ca/support-and-resources-for-lawyers/act-rules-and-code/code-of-professional-conduct-for-british-columbia/.

[20] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What Does it Mean for a Language Model to Preserve Privacy? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 2022.

[21] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901, 2020.

[22] Dana Burchardt. Backlash against the Court of Justice of the EU? The Recent Jurisprudence of the German Constitutional Court on EU Fundamental Rights as a Standard of Review. *German Law Journal*, 21:1–18, 2020. doi: doi:10.1017/glj.2020.16.

[23] Naomi R Cahn. Looseness of legal language: The reasonable woman standard in theory and in practice. *Cornell L. Rev.*, 77:1398, 1991.

[24] Canadian Judicial Council. Ethical Principles for Judges, 2004. https://cjc-ccm.ca/cmslib/general/news_pub_judicialconduct_Principles_en.pdf.

[25] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.

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

[27] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

[28] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy, 2019. Association for Computational Linguistics.

[29] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, 2020.

[30] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, 2022.

[31] Sumi Cho. Post-racialism. *Iowa L. Rev.*, 94:1589, 2008.

 https://doi.org/10.52202/068431-2119

[32] Maximin Coavoux, Shashi Narayan, and Shay B Cohen. Privacy-preserving Neural Representations of Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, 2018.

[33] I Glenn Cohen and Michelle M Mello. HIPAA and protecting health information in the 21st century. *JAMA*, 320(3):231–232, 2018.

[34] Elaine Craig. The ethical obligations of defence counsel in sexual assault cases. *Osgoode Hall L. J.*, 51:427, 2013.

[35] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

[36] Department of Veterans Affairs. Health Insurance Portability and Accountability Act Applicability in VBA, 2003. https://www.va.gov/ogc/docs/ADV3-2003.pdf.

[37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

[38] Jana DiCosmo. Racism in the Legal Profession: A Racist Lawyer Is an Incompetent Lawyer. *Nat'l Law. Guild Rev.*, 75:82, 2018.

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

[40] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, 2021.

[41] Zachary Elkins, Tom Ginsburg, James Melton, Robert Shaffer, Juan F Sequeda, and Daniel P Miranker. Constitute: The world's constitutions to read, search, and compare. *Journal of web semantics*, 27:10–18, 2014.

[42] Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. Research challenges in designing differentially private text generation mechanisms. *arXiv preprint arXiv:2012.05403*, 2020.

[43] Richard Thompson Ford. Racial epithets and racial etiquette. *Capital University Law Review*, 49(4):527–534, 2021.

[44] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[45] Alan E Garfield. To swear or not to swear: using foul language during a Supreme Court oral argument. *Wash. U. L. Rev.*, 90:279, 2012.

[46] Jamal Greene. The anticanon. *Harv. L. Rev.*, 125:379, 2011.

[47] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. All You Need is "Love" Evading Hate Speech Detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12, 2018.

[48] Suchin Gururangan, Dallas Card, Sarah K Drier, Emily K Gade, Leroy Z Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A Smith. Whose language counts as high quality? Measuring language ideologies in text data selection. *arXiv preprint arXiv:2201.10474*, 2022.

[49] Laura Hanu and Unitary team. Detoxify. Github, 2020. https://github.com/unitaryai/detoxify.

[50] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[51] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

[52] David Hausman, Daniel E Ho, Mark Krass, and Anne M McDonough. Executive Control of Agency Adjudication: Capacity, Selection and Precedential Rulemaking. *Journal of Law, Economics & Organization*, 40, 2024.

[53] Allison Hegel, Marina Shah, Genevieve Peaslee, Brendan Roof, and Emad Elwany. The Law of Large Documents: Understanding the Structure of Legal Contracts Using Visual Cues. *arXiv preprint arXiv:2107.08128*, 2021.

[54] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

[55] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. *arXiv preprint arXiv:2103.06268*, 2021.

[56] Zihan Huang, Charles Low, Mengqiu Teng, Hongyi Zhang, Daniel E Ho, Mark S Krass, and Matthias Grabmair. Context-aware legal citation recommendation using deep learning. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 79–88, 2021.

[57] Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE, 2020.

[58] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Somaieh Nikpoor, Jörg Frohberg, Aaron Gokaslan, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. Data governance in the age of large-scale data-driven language technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2206–2222, 2022.

[59] Jigsaw and Google's Counter Abuse Technology team. Perspective, 2021. https://perspectiveapi.com/.

[60] Sheri Lynn Johnson, John H Blume, and Patrick M Wilson. Racial Epithets in the Criminal Process. *Mich. St. L. Rev.*, page 755, 2011.

[61] Lisa LaPlant Jon Quandt, Eric Mill. govinfo. Github. https://github.com/usgpo/api, 2018.

[62] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. *arXiv preprint arXiv:2202.06539*, 2022.

[63] Randall Kennedy and Eugene Volokh. The new taboo: Quoting epithets in the classroom and beyond. *Cap. U. L. Rev.*, 49:1, 2021.

[64] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit*, pages 79–86, 2005.

[65] Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. ADePT: Auto-encoder based Differentially Private Text Transformation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2435–2439, 2021.

 https://doi.org/10.52202/068431-2119

[66] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

[67] David S Law and Mila Versteeg. The declining influence of the United States Constitution. *N. Y. U. L. Rev.*, 87:762, 2012.

[68] Law Society of Ontario. Rules of Professional Conduct, 2022. https://lso.ca/about-lso/legislation-rules/rules-of-professional-conduct.

[69] Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. Do language models plagiarize? *arXiv preprint arXiv:2203.07618*, 2022.

[70] Alyssa Lees, Daniel Borkan, Ian Kivlichan, Jorge Nario, and Tesh Goyal. Capturing covertly toxic speech via crowdsourcing. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 14–20, Online, April 2021. Association for Computational Linguistics.

[71] Spyretta Leivaditi, Julien Rossi, and Evangelos Kanoulas. A benchmark for lease contract review. *arXiv preprint arXiv:2010.10386*, 2020.

[72] Mark A Lemley and Bryan Casey. Fair learning. *Tex. L. Rev.*, 99:743, 2020.

[73] Justin D Levinson, Mark W Bennett, and Koichi Hioki. Judging implicit bias: A national empirical study of judicial sterotypes. *Fla. L. Rev.*, 69:63, 2017.

[74] Yitong Li, Timothy Baldwin, and Trevor Cohn. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, 2018.

[75] Benjamin L Liebman, Margaret E Roberts, Rachel E Stern, and Alice Z Wang. Mass Digitization of Chinese Court Decisions. *Journal of Law and Courts*, Fall:176–201, 2020.

[76] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139, 2019.

[77] Michael Lissner and Brian W Carver. Courtlistener. com: A platform for researching and staying abreast of the latest law. 2010. http://www.courtlistener.com.

[78] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[79] Eneldo Loza Mencía and Johannes Fürnkranz. Efficient multilabel classification algorithms for large-scale problems in the legal domain. In *Semantic Processing of Legal Texts*, pages 192–215. Springer, 2010.

[80] Lingjuan Lyu, Xuanli He, and Yitong Li. Differentially private representation for nlp: Formal guarantee and an empirical study on privacy and fairness. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2355–2365, 2020.

[81] Sibella Matthews, Vincent Schiraldi, and Lael Chester. Youth justice in Europe: Experience of Germany, the Netherlands, and Croatia in providing developmentally appropriate responses to emerging adults in the criminal justice system. *Justice Evaluation Journal*, 1(1):59–81, 2018.

[82] H Brendan McMahan, Galen Andrew, Ulfar Erlingsson, Steve Chien, Ilya Mironov, Nicolas Papernot, and Peter Kairouz. A general approach to adding differential privacy to iterative training procedures. *arXiv preprint arXiv:1812.06210*, 2018.

[83] Microsoft. Microsoft presidio. Github., 2022. https://github.com/microsoft/presidio/.

[84] Eric Mill. Opening up government reports through teamwork and open data. *OpenGov Voices*, 2014.

[85] Suzanne J Miller. Judicial language in new jersey sexual violence cases. *Rutgers U. L. Rev.*, 73:141, 2020.

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

[87] Nancy Morawetz. A better balance for federal rules governing public access to appeal records in immigration cases. *Hastings LJ*, 69:1271, 2017.

[88] Deborah A Morgan. Not gay enough for the government: Racial and sexual stereotypes in sexual orientation asylum cases. *Law & Sexuality: Rev. Lesbian, Gay, Bisexual & Transgender Legal Issues*, 15:135, 2006.

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

[90] Beau Norgeot, Kathleen Muenzen, Thomas A Peterson, Xuancheng Fan, Benjamin S Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, et al. Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes. *NPJ digital medicine*, 3(1):1–8, 2020.

[91] Nicholas M Pace and Julia Pollak. Provider fraud in california workers' compensation: Selected issues, 2017. https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1703/RAND_RR1703.pdf.

[92] Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States, July 2022. Association for Computational Linguistics.

[93] Prakash Poudyal, Jaromír Šavelka, Aagje Ieven, Marie Francine Moens, Teresa Gonçalves, and Paulo Quaresma. Echr: legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, 2020.

[94] Praatika Prasad. Implicit racial biases in prosecutorial summations: Proposing an integrated response. *Fordham L. Rev.*, 86:3091, 2017.

[95] Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. Deconfounded lexicon induction for interpretable social science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1615–1625, 2018.

[96] National Historical Publications and Records Commission. Founders online, 2010. https://founders.archives.gov/.

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

 https://doi.org/10.52202/068431-2119

Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

[98] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[99] Abigail M Reecer. The ethical dilemmas of the office of legal counsel in the wake of a whistleblower complaint. *Geo. J. Legal Ethics*, 33:769, 2020.

[100] Douglas Rice, Jesse H Rhodes, and Tatishe Nteta. Racial bias in legal language. *Research & Politics*, 6(2):2053168019848930, 2019.

[101] Federico Ruggeri, Francesca Lagioia, Marco Lippi, and Paolo Torroni. Detecting and explaining unfairness in consumer contracts through memory networks. *Artificial Intelligence and Law*, pages 1–34, 2021.

[102] Steven Ruggles, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. Differential privacy and census data: Implications for social and economic research. In *AEA papers and proceedings*, volume 109, pages 403–08, 2019.

[103] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. *arXiv preprint arXiv:1910.14659*, 2019.

[104] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[105] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[106] Arman Sarvarian. Common ethical standards for counsel before the european court of justice and european court of human rights. *European Journal of International Law*, 23(4):991–1014, 2012.

[107] Eran Shalev. Ancient masks, american fathers: Classical pseudonyms during the american revolution and early republic. *Journal of the Early Republic*, 23(2):151–172, 2003.

[108] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1310–1321, 2015.

[109] Amy Sholsberg, Evan Mandery, and Valerie West. The expungement myth. *Albany Law Review*, 75:1229–1242, 2011.

[110] Social Security Administration. HIPAA and the Social Security Disability Programs. https://www.ssa.gov/disability/professionals/hipaa-cefactsheet.htm.

[111] Congzheng Song and Ananth Raghunathan. *Information Leakage in Embedding Models*, page 377–390. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450370899.

[112] Harold J Spaeth, Lee Epstein, Andrew D Martin, Jeffrey A Segal, Theodore J Ruger, and Sara C Benesh. Supreme court database, version 2013 release 01. *Database at http://supremecourtdatabase.org*, 2013.

[113] Mike Spector, Jaimi Dowdell, and Benjamin Lesser. How secrecy in U.S. courts hobbles the regulators meant to protect the public. *Reuters*, 2010. https://www.reuters.com/investigates/special-report/usa-courts-secrecy-regulators.

[114] Dan Sperber and Deirdre Wilson. Irony and the use-mention distinction. *Philosophy*, 3: 143–184, 1981.

[115] Supreme Court of Canada. Sherman Estate v. Donovan, 2021. https://www.canlii.org/en/ca/scc/doc/2021/2021scc25/2021scc25.html.

[116] The President and Fellows of Harvard University. Caselaw access project. https://case.law/api/.

[117] Mark V Tushnet. *The Warren Court in historical and political perspective*. University of Virginia Press, 1993.

[118] U.K. Judicial Conduct Investigations Office. *Annual Report 2015–2016*. 2016. https://jciodev.microsoftcrmportals.com/_entity/annotation/61785fbb-752a-eb11-a813-000d3a0bacd3.

[119] United Kingdom Bar Standards Board. Bar Standards Board Handbook. https://www.barstandardsboard.org.uk/the-bsb-handbook.html?part=E3FF76D3-9538-4B97-94C02111664E5709&audience=&q=.

[120] United Nations Office on Drugs and Crime. *Commentary on the Bangalore Principles of Judicial Conduct*. 2007. https://www.unodc.org/documents/nigeria/publications/Otherpublications/Commentry_on_the_Bangalore_principles_of_Judicial_Conduct.pdf.

[121] United States Court of Appeals for the Second Circuit. Sealed Plaintiff v. Sealed Defendant, 2008.

[122] U.S. Election Assistance Commission. Availability of state voter file and confidential information, 2020. https://www.eac.gov/sites/default/files/voters/Available_Voter_File_Information.pdf.

[123] Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, 2018.

[124] Eugene Volokh. Pseudonymous litigation, 2021. https://www2.law.ucla.edu/volokh/pseudonym.pdf.

[125] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, 2020.

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

[127] Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman SM Chow. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, 2021.

[128] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*, 2022.

[129] Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 159–168, 2021.

[130] Victor Zhou. profanity-check. Github. https://github.com/vzhou842/profanity-check, 2019.

 https://doi.org/10.52202/068431-2119

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see our ethics statement.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Please see Appendix A, where we provide a point-by-point discussion of how our paper conforms to the 2022 NeurIPS ethics guidelines.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix D.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix C.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Please see Appendix A, where we provide a point-by-point discussion of how our paper conforms to the 2022 NeurIPS ethics guidelines.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

https://doi.org/10.52202/068431-2119