MOTPP: A Multi-Genre Corpus of Online Terms and Privacy Policies

Anonymous ACL submission

Abstract

Legal notices are pervasive online. Digital spaces are littered with legally binding terms and policies that govern digital rights and shape access to justice. Yet many of those texts are opaque — difficult to comprehend and study. Our research addresses that gap. First, we introduce the Multi-Genre Online Terms and Privacy Policies (MOTPP), a synchronic dataset composed of the online terms and privacy policies of prominent digital platforms across nine genres. The dataset contains 835 texts and 5.89 million tokens. Second, we provide an interdisciplinary analysis that illustrates linguistic features of the corpus and presents machine learning tools for scrutinizing digital contracts at scale. Our exploratory application leverages machine learning and synthetic data to analyze key content for consumers, focusing on terms that determine access to justice. The annotated dataset, models, and other resources for this paper are available at GitHub and Hugging Face.

1 Introduction

Doctor Kanokporn Tangsuan was a family medicine specialist in New York. While vacationing in Florida, she went to an Irish pub at a Walt Disney World resort for lunch with her husband, Jeffrey Piccolo, and his mother. Knowing that Doctor Tangsuan had life-threatening food allergies, they took extensive precautions when ordering from the menu. Despite that, she had an acute anaphylactic reaction to the food after leaving the restaurant. Sadly, she died soon afterwards, aged 42.

When her family sued the pub and Disney in a Florida court, an obscure legal term in an equally obscure contract suddenly entered the spotlight. In 2019, about five years before their lunch at the pub, Mr. Piccolo signed up for a Disney+ account online (**?**). The account registration terms contained a broad arbitration clause. According to Disney's lawyers, that clause blocked the family from suing the company in court. They could only bring legal claims via private arbitration.

The incident attracted media attention and public scorn, which ultimately compelled Disney to change their stance on arbitration in this particular dispute. Still, the events highlighted a systemic trend: the prevalence and reach of arbitration clauses in online contracts. Doctor Tangsuan's situation is far from an isolated case. Whether people encounter racial discrimination on Airbnb (**?**) or suffer life-changing injuries on an Uber ride (**?**), arbitration clauses determine access to justice at societal scale, particularly in the United States.

Almost all digital applications and websites impose standard form contracts on their users. These texts are often referred to as terms of use, terms of service, and user agreements (collectively, TOUs). As binding contracts, TOUs are the central legal construct between users and digital platforms. Like most contracts, TOUs define rights and responsibilities, allocate risk, set mechanisms for handling disputes, specify governing law, and more. Most websites also have a privacy policy (PP), which provides notice to users about data collection and management. A persistent challenge for the public is the opaqueness of these documents, which are lengthy and linguistically complex. Despite their far-reaching economic and social implications, navigating TOUs and PPs is virtually impossible for the general public.

This paper introduces the Multi-Genre Online Terms and Privacy Policies (MOTPP) corpus, a synchronic dataset composed of 421 TOUs and 414 PPs from nine genres of digital platforms. MOTPP contains approximately 5.89 million tokens. The dataset is unlike any other widely available legal language dataset in terms of the scope, genre classifications of the texts, and the annotations for key legal terms. Following FAIR principles (Wilkinson et al., 2016), it is freely available at GitHub and

Hugging Face.

In addition to the MOTPP corpus, this paper introduces frameworks that facilitate the analysis of these contracts at scale. We demonstrate a workflow for content detection, which enables users to efficiently navigate this complex legal landscape with minimal manual inputs. In our case study, we propose several strategies for generating synthetic training data as we evaluate domain-specific classifiers that detect arbitration clauses, opt-out provisions, and class waivers.

The contributions of this interdisciplinary work include: (a) a novel and publicly available dataset of legal texts that govern digital rights and privacy, (b) a baseline analysis of linguistic characteristics across TOUs and PPs from various genres of digital platforms, (c) an exploratory application of an automated classifier that detects and extracts key content in TOUs, and (d) code, synthetic data, and models to reproduce the results. These contributions offer public resources for researchers, policymakers, citizens, and communicators seeking to navigate this realm of digital governance.

2 Background

2.1 Disclosure in Digital Spaces

Disclosure is central to "notice and choice" and "informed consent" models of digital governance and privacy. In theory, such models enable individuals to self-manage their digital rights (Solove, 2013). In practice, people are inundated with a never-ending stream of notices and privacy decisions. Notices are ubiquitous in digital spaces: cookie banners, pop-up notices, privacy policies, contract terms, updates, and so on.

Because digital platforms mediate unprecedented amounts of data and human activity, their legal texts play an outsized role in digital governance (Kim and Telman, 2015; **?**). But most of those texts are incomprehensible to the general public. Compared to other forms of written language, they are exceptionally dense and complex in their linguistic forms. For a variety of reasons, very few people ever attempt to read them (Bakos et al., 2014).

2.2 Legal Clauses of Interest

For our initial experiments, we focused on dispute resolution terms: arbitration clauses, opt-out provisions, and class waivers. We selected those clauses for experimentation because of their importance for digital rights and access to justice.

Arbitration is a form of private dispute resolution. As an alternative to litigation, arbitration takes place outside of court systems. Like litigation, arbitration is adversarial and binding. Unlike litigation, arbitration is private and often confidential. Arbitration also lacks fundamental features of judicial proceedings: juries and the right to appeal. Procedures and discovery are also streamlined in arbitration, which can be faster and cheaper than litigation.

Because they curb the public's right to access justice, arbitration clauses are a controversial feature of consumer contracts. They are are especially prevalent in the United States, where the legal system enables companies to funnel consumer disputes toward arbitration. **Opt-out** provisions within arbitration clauses typically offer users a limited window to notify the platform of their preference to opt-out of arbitration. **Class waivers**, commonly paired with arbitration clauses, prevent users from bringing claims against the platform as a class or a group.

3 Related Work

Previous studies have assessed the linguistic complexity of legal texts, including online terms and policies. Early contributions assessed the reading difficulty of online TOUs using traditional readability formulas like the Flesch-Kincaid test (Rustad and Koenig, 2014; Benoliel and Becher, 2019). More recent works have broadened the scope of analysis with more robust linguistic metrics. Martínez et al. (2022) measured center-embedded clauses and passives in contract language. In addition, previous work measured the syntactic complexity of verb and noun structures in a dataset of TOUs and PPs.

There are also a number of related datasets. For instance, Amos et al. (2021) and **?** assembled large PP datasets, each with over one million policies. Wagner (2022) collected a corpus of PPs from 1996-2021 and assessed aspects of longitudinal change. Other datasets focus on TOUs. Marotta-Wurgler and Taylor (2013) examine change over time in standard form consumer agreements.

Other collections include targeted datasets with annotations, such as a corpus of 510 contracts in twenty-five different categories (Hendrycks et al., 2021) or a corpus of German consumer contracts (**?**). Researchers are also developing tools to navigate TOUs and PPs. Claudette, an automated detec-

tor of potentially unfair clauses in online consumer contracts (Lippi et al., 2019). **?** designed a browser extension to detect opt-out choices in privacy policies. Studies such as Lippi et al. (2019) are increasingly testing machine learning and natural language processing techniques to analyze and assess contract language. Such techniques are emerging in interdisciplinary legal NLP research more broadly (Choi, 2023).

Despite these valuable contributions, gaps in publicly available data and toolkits remain. MOTPP offers a dataset of TOUs and PPs from prominent platforms across nine genres with annotations of key terms. We expand on that contribution by illustrating the linguistic characteristics of these legal texts. For context, we compare those characteristics across document categories (TOUs versus PPs), genres (fintech versus social), and external corpora (contract language versus general English).

Finally, we test a machine learning workflow for content exploration related to disputes and access to justice. We explore the use of synthetic texts as training data for specialized, domain-specific classifiers by evaluating their ability to detect complex (legal) concepts in the TOUs (**?**). Synthetic data has been particularly popular for training and fine-tuning language models (**?**), but recent studies highlight mixed results in the context of computational and social sciences research (**?**). In this paper, we follow and build on **?** who propose zero and few-shot strategies for generation. We expand on their work by introducing a "contrastive few-shot" prompt. Our strategy aims to enhance variety in the generated data by instructing the model to build on randomly sampled combinations of positive or negative examples.

4 Corpus Compilation, Composition, Statistics, and Metadata

Our dataset is synchronic and consists of 421 TOUs and 414 PPs from digital platforms. Each policy was manually scraped from the platform's official website and followed the same set of procedures for processing. The workflow proceeded as follows. Scraped texts were first saved as document files, then converted through a shell script to plain text. The texts were then cleaned with a python script to remove irrelevant characters and aberrations in formatting.

Table 1 illustrates the corpus composition across

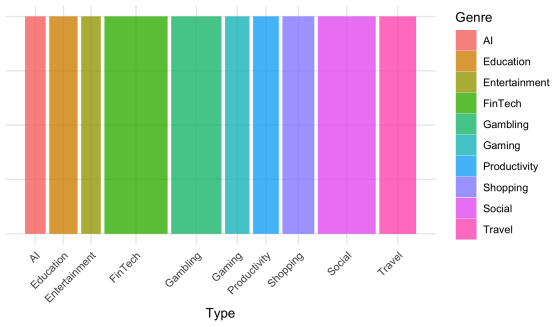

Figure 1: Distribution of Tokens per Contract Genre

the nine specific genres of platforms. Table 1 shows the number of TOUs, PPs, and the respective token counts by category. The aggregate word counts show that TOUs tend to be longer than PPs, averaging 6,642 and 4,009, respectively.

Figure 1 visualizes the aggregate (TOUs plus PPs) tokens by genre. The social genre is the largest category in the dataset, followed by fintech and gambling. Within the social genre, there are distinguishable sub-genres: chat, creator, social network, Q&A, subculture, alt-tech, and others (Zuckerman and Rajenda-Nicolucci, 2021). For this study, we include these sub-genres as well as dating applications within the social category.

Other platform genres include education, entertainment, productivity, shopping, gaming, and AI. These categories are generally consistent with application categories on the Apple App Store and the Google Play app store. Platforms were then selected within our categories based on download rankings as well as popularity data from Statista and the Pew Research Center (Clement, 2023).

5 Linguistic Characteristics and Example Analyses

In this section, we detail linguistic characteristics of the corpus, including annotations for part-of-speech (POS) tags and linguistic complexity metrics. Our processing for linguistic characteristics and complexity is conducted in R using quanteda (Benoit et al., 2018), cleanNLP (Arnold, 2021), udpipe (Wijffels, 2023), and the tidyverse (Wickham et al., 2019).

Tables 2 and 3 show the normalized results for verbs, nouns, embedding, and center-embedding. The results are divided by policy type, genre, and POS. Results for TOUs are in Table 2 and PPs are in Table 3.

The density of verbs and nouns is normalized as

Genre	TOUs	Tokens (TOUs)	PPs	Tokens (PPs)
Travel	28	280929	28	244335
Social Media	75	504107	74	324320
Education	39	220339	38	182830
Entertainment	23	155400	23	122449
FinTech	75	625368	71	279753
Gambling	47	471790	47	247774
Productivity	38	212868	38	153715
Shopping	31	293305	31	160786
Gaming	33	221486	33	124671
AI	32	159871	31	132463
Totals	421	3630121	414	2258205

Table 1: Corpus Composition: TOUs and PPs

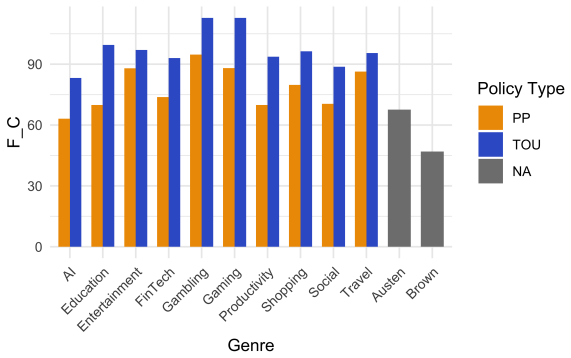

Figure 2: Verbal Complexity Results by Genre

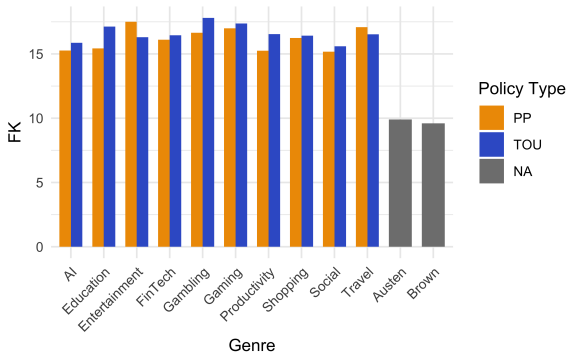

Figure 3: Flesch-Kincaid Results by Genre

a function of total words per contract category per five-hundred words, while embedding is normalized at the sentence level. Similar to Martínez et al. (2022), we counted embedded clauses as those containing tokens tagged by udpipe as having clausal subject, clausal complement, open clausal complement, adjectival clause, and adverbial clause Universal Dependency relationships (Wijffels, 2023). Clauses were considered center-embedded if tokens tagged with these relationships were not followed by punctuation (Martínez et al., 2022).

We also use the POS tags to measure the syntactic complexity of verb structures. For that measurement, we apply Fichtner's C (F_C), which characterizes verbal complexity as a result of lexical verbs per sentence (Fichtner, 1980; Mollet et al., 2010; Gries, 2016). Higher F_C scores indicate more complex verb structures in a text.

Figure 2 shows the average verbal complexity across the genres in MOTPP and compares those results with two prominent external corpora: the *Brown Corpus of General American English* (Francis and Kučera, 1979) and the *Collected Works of Jane Austen* (Silge and Robinson, 2016). Interestingly, verbal complexity is higher for TOUs for all document categories — above privacy policies, the Austen Corpus, and the Brown Corpus. Across genres, gaming TOUs have the highest average verbal complexity overall, following by gambling, education, and shopping.

Figure 3 displays the results of a traditional readability metric, the Flesch-Kincaid (F-K) score. Like other traditional readability metrics, F-K approximates reading difficulty as a function of words per sentence (representing syntactic complexity) and syllables per word (representing lexical difficulty) (Kincaid et al., 1975). As metrics for scoring reading difficulty, traditional readability metrics are fundamentally flawed; they have major technical and theoretical shortcomings. Despite those limitations, we favor calculating the F-K scores to enable comparison with previous legal and interdisciplinary studies on contract language. The results shown in Figure 3 suggest that TOUs and PPs are much more difficult to comprehend than a sample of literature (the Austen Corpus) and general English (the Brown Corpus).

The results in Figures 2 and 3 also reveal an interesting divergence between the metrics. Whereas the F_C results highlight a marked difference in syntactic complexity between TOUs and PPs, the F-K results do not. As the F-K results show, PPs

Genre	Verbs	Nouns	Embedding	Center-Embedding
AI	51.3	138.8	1.5	0.61
Education	49.1	141.4	1.6	0.63
Entertainment	49.1	140.1	1.6	0.62
FinTech	51.0	146.6	1.6	0.63
Gambling	50.2	137.2	1.7	0.65
Gaming	48.8	139.8	1.7	0.64
Productivity	47.2	139.2	1.6	0.60
Shopping	48.2	141.1	1.6	0.63
Social	49.3	137.3	1.5	0.59

Table 2: Verb, Noun Density and Embedding: TOUs

Genre	Verbs	Nouns	Embedding	Center-Embedding
AI	57.3	141.1	1.3	0.58
Education	49.1	141.4	1.6	0.63
Entertainment	58.4	148.5	1.6	0.63
FinTech	56.5	145.9	1.6	0.63
Gambling	56.6	139.5	1.75	0.66
Gaming	55.6	139.5	1.67	0.64
Productivity	57.2	140.4	1.5	0.59
Shopping	55.1	143.3	1.4	0.63
Social	58.1	135.7	1.6	0.63

Table 3: Verb, Noun Density and Embedding: PPs

and TOUs contain relatively long sentences and long words relative to the Brown and Austen corpora. However, variations in syntactic complexity between TOUs and PPs go undetected by F-K. This divergence is consistent with previous studies that compare more nuanced features of syntactic complexity in TOUs and PPs (Samples et al., 2024). We suspect that this divergence highlights the lower sensitivity — or perhaps dullness — of traditional readability metrics to features of linguistic complexity.

6 Navigating the Legal Landscape

In this section, we discuss a machine learning workflow for the exploration of the legal content in MOTPP. While crucially important to our online as well as offline lives, legal documents remain opaque, difficult to access, and bewildering in their complexity. The workflow we envisage allows users to analyze key TOU content at scale, leveraging machine learning to obtain a corpus-level overview of content distribution. From a technical perspective, our experiments focus on strategies that effectively detect content based on only a minimal amount of human input in the form of "seed examples" and annotations. Toward this end, we propose and evaluate a machine learning pipeline that leverages synthetic data to facilitate the exploration of textual data, allowing for more critical engagement with these TOUs. Our ultimate goal is to democratize access to these often-overlooked but important legal texts through machine learning.

In this paper, we report on the initial efforts towards building an analytical space that enables specialists and non-specialists alike to demystify and explore the content of the TOUs and PPs they encounter. This section proceeds as follows. We first introduce the types of information we wish to extract and then turn to describing the input data. Then, we elaborate on the various techniques used for generating synthetic data and evaluate the performance of models trained on such artificially generated examples. Lastly, we apply the best strategy to our corpus, to get a sense how well these models perform "in the wild."

6.1 Machine Learning Workflow

From a machine learning point of view, this paper is principally concerned with evaluating the role of synthetic data for training specialized classifiers based on minimal human input. Whereas human input is essential to steer the model in the right direction, we want to test the effectiveness of synthetic data in the direction signaled by the user to build more capable domain-specific and open-source models for the exploration of legal texts.

Before we turn to outlining the workflow and discussing the results, we address a common query: Why not simply use ChatGPT for document classification? We do compare our results to GPT-4o, which works well. However, while we acknowledge that large language models perform

well at this classification task, we are more interested in exploring how we can inject knowledge from the larger models into smaller, specialized models through synthetic data. This will enable us to operate faster, cheaper, and hopefully with similar efficacy as the mammoth models.

6.1.1 Seed examples and annotations

We start with set text fragments, also referred to as "seeds," S_t for a target category t. This can be a very small number. For our experiments, we used a handful of arbitration, opt-out and class waiver clauses as the initial inputs. We expand these seeds to a larger set of retrieved examples R_t from the corpus C.

By embedding seeds s_{seed} in S_t each sentence s_c in C we can compute the similarity of the input examples to content in the corpus. For our purposes, we encoded each text using the Nomic AI Text Embedding Model (Brown, 2020), and then selected 25 examples (from C) that exhibited high similarity to these seeds for further annotation A_t. We annotated these text fragments as either belonging to t or not.[1]

6.1.2 Synthetic data generation

We leveraged these annotated snippets to steer the synthetic examples Syn_t in a particular semantic direction, which hopefully captures our clauses (or concept of interest t) with accuracy. Below, we evaluate both the efficacy of different prompting strategies for generating synthetic data, as well as the number of examples needed to build an adequate domain-specific classifier c_t In total, we generate 250 positive and negative synthetic examples for each clause type, relying on different prompting strategies. The exact prompt templates with examples can be found in the Appendix, but here we provide an overview of the most important characteristics. The prompts for generating positive examples start with: "You are a helpful AI that generates a new example a t clause." With t being either arbitration, opt-out, or class waiver.

- **Zero-shot**: We provide a definition of t and ask the model to generate a new example.
- **Few-shot**: We provide a definition and three positive examples randomly sampled from At. We then instruct the model to generate a new example of t, with additional directions: "You can combine elements of the three examples, but have to change the word order, use synonyms, and change the sentence structure. The end result, however, has to remain a t clause from a legal and semantic point of view."
- **Few-shot and contrastive**: We repeat the few-shot prompt, but add three randomly sampled negative or "contrastive" examples. We add to the instruction: "Make sure the new example is very different from the contrastive examples. The end result, however, has to remain a t clause from a legal and semantic point of view."

6.1.3 Model training and evaluation

To evaluate different approaches to data generation, we manually annotated 100 examples as a test set.[2] We fine-tuned a distilbert-base-uncased model on a concatenation of A_t and Syn_t for 5 epochs with a learning rate of 1e-5, using AdamW as an optimizer. Given the very small amount of training data, fine-tuning larger models did not make sense, and in the few experiments we conducted, results were equal if not worse. Our initial experiments include only annotated and synthetic data. To assess how the model would fare in a more realistic scenario, where it would encounter other types of contract language, we added randomly sampled sentences C as negative examples to both the training set (n=200) and the test set (n=50). This step injects more variety and noise in the evaluation procedure and mimics how the model might fare in such an environment.[3] In each of our experiments we report results for these different training routines.

Our experiments primarily gauge the relative improvements of adding synthetic data to a small set of 25 annotated examples which are used to train our baseline model. While small, this model could still be competitive, as a few words tend to contain strong lexical signals for these clause types. As a "skyline" of sorts we also report the results obtained using GPT-4o. We report the best

[1] More technically, we created two matrices, one which encoded the sentences M_C while the other M_S comprised the vector representation of the seeds. We then computed the cosine similarity between these two matrices, which resulted in a new matrix M_{sim} in which each sentence in C is scored with respect to all the examples in S_t. Subsequently, we sorted M_{sim} by the maximum value in each row and sampled 25 examples among the top 500 most similar sentences for annotation.

[2] Following the same sample strategy as explained in the previous footnote.

[3] Of course, the automatically assigned negative class might be incorrect, but the chances are small. Moreover, we are most interested in relative improvements.

accuracy and macro-f1 scores after training for 5 epochs.

6.1.4 Results

Table 4 reports the accuracy and macro-f1 scores for different prompting and data generation strategies. In all cases, the generation of synthetic data, even a few examples, resulted in better training data. Overall, the right instructions and a handful of meaningful examples substantially improved model performance. These steps enabled the model to find occurrences of clauses that might have otherwise escaped one trained solely on examples extracted from contracts. Overall, the few-shot scenarios delivered the best results, greatly improving the macro-f1 scores (e.g., from 0.45 to 0.95 for class waivers). Based on these experiments, we cannot yet conclude that contrastive prompting consistently out-runs other approaches, but it appears to be a strong competitor. Improvements are not evenly spread across all clause types, with the opt-out clause posing more challenges, even though we still observe an increase of 0.31 points. Given the small size of the model and the minimal amount of training, the gains are impressive and are gradually coming closer to GPT-4o, but a consistent gap remains in all scenarios. In future research, we aim to investigate the results of increasing the diversity of synthetic data and of generating more data.

As a case study and additional evaluation, we analyze the TOUs of AI applications. For this subset, we have access to document-level annotations (as opposed to sentence-level labels used in the previous evaluations) which we can compare against our different fine-tuned models. In this case we would label the complete contracts as either zero or one, based on whether it contained a specific clause type or not.

Table 6 reports the extent to which the models corresponded to human annotations at the contract level. We increase the size of the training and report macro-f1 scores for each scenario. For example, the *train set only* with n equal to 25, indicates that we fine-tuned our models on this number of annotated examples. For the *few shot* scenario we added 50 synthetic texts (25 positive and negative examples) to the training set.

In almost all scenarios, this addition of synthetic data results in better scores, showing potential benefits. Still, at the same time, creating synthetic data has yet to emerge as a panacea that will enable us solve the problem of annotation scarcity. Only after training on all annotated and synthetic data does the macro-f1 score indicate that the classifier works with increasing reliability.

7 Limitations

Several limitations are worth noting. First, while these data represent a significant number of prominent digital platforms across various genres, they are not necessarily representative of the entire population. Second, our focus on English-language texts influences the results, as contracts differ across languages and jurisdictions. Third, the dataset is synchronic, yet TOUs and PPs have shown to be quite plastic, changing frequently (authors). Fourth, no single approach to reading difficulty is exhaustive. Our approach to linguistic complexity captures an important spectrum of characteristics but does not cover every dimension of complexity in these texts (**?**). Finally, our experiment with the machine learning classifier is limited to arbitration-related clauses and a subset of the data (two of the nine genres). While arbitration-related clauses are key for access to justice in the consumer context, we acknowledge that there are other terms of interest we have yet to classify.

8 Conclusion

This interdisciplinary work introduces MOTPP, a novel corpus of digital legal texts. With annotations of key terms, MOTPP is a curated dataset that represents TOUs and PPs across nine genres of digital platforms. In addition to offering linguistic analyses and potential avenues for continued research, this paper pilots a toolkit: a machine learning classifier that identifies and extracts terms related to disputes and access to justice. This work offers publicly available resources for legal scholars and NLP practitioners, policymakers, and citizens alike.

9 Statement of Reproducibility

Our dataset and code are available at GitHub and Hugging Face so that all of our methods and analyses can be reproduced.

References

Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayar. 2021. Privacy policies over time: Curation and analysis of a million-document dataset. *IW3C2 (International World Wide Web Conference Committee).*

			accuracy			macro-f1
clause prompt	arbitration	class waiver	opt-out	arbitration	class waiver	opt-out
train set only	0.38	0.51	0.60	0.30	0.34	0.38
zero shot 50	0.82	0.80	0.61	0.80	0.80	0.44
zero shot 100	0.85	0.81	0.66	0.84	0.81	0.56
zero shot 250	0.84	0.80	0.70	**0.85**	0.80	0.63
few shot 50	0.84	0.83	0.67	0.83	0.83	0.56
few shot 100	0.84	0.85	0.73	0.83	0.85	0.66
few shot 250	0.84	0.92	0.71	0.83	**0.92**	0.65
contrastive few shot 50	0.75	0.88	0.65	0.75	0.88	0.51
contrastive few shot 100	0.84	0.91	0.69	0.83	0.91	0.60
contrastive few shot 250	0.85	0.92	0.76	**0.85**	0.91	**0.71**
gpt-4o	0.97	0.99	0.95	0.95	0.99	0.95

Table 4: Accuracy and macro-f1 scores for different synthetic data generation strategies

			accuracy			macro-f1
clause prompt	arbitration	class waiver	opt-out	arbitration	class waiver	opt-out
train set only	0.77	0.81	0.85	0.43	0.45	0.46
zero-shot 100	0.84	0.94	0.86	0.71	0.90	0.48
zero-shot 200	0.87	0.94	0.87	0.79	0.89	0.57
zero-shot 500	0.90	0.96	0.88	0.85	0.94	0.66
few-shot 100	0.88	0.95	0.85	0.81	0.90	0.46
few-shot 200	0.92	0.96	0.90	0.88	0.93	0.71
few-shot 500	0.93	0.97	0.90	**0.91**	**0.95**	0.71
contrastive few-shot 10	0.85	0.94	0.86	0.73	0.90	0.51
contrastive few-shot 200	0.89	0.96	0.88	0.81	0.94	0.66
contrastive few-shot 500	0.92	0.97	0.91	0.88	**0.95**	**0.77**
gpt-4o	0.97	0.99	0.95	0.95	0.99	0.95

Table 5: Accuracy and macro-f1 scores for different synthetic data generation strategies, with randomly added negative observations to both train and test

prompt	n	arbitration	opt-out	class waiver
train set only	25	0.29	0.37	0.32
train set only	50	0.66	0.37	0.32
train set only	100	0.71	0.37	0.65
few shot	25	0.37	0.29	0.46
few shot	50	0.71	0.40	0.61
few shot	100	**0.83**	**0.58**	**0.73**

Table 6: Macro-f1 scores comparing the classifier outputs (for training data of different size) to manual document level annotations

Taylor Arnold. 2021. A tidy data model for natural language processing using cleannlp. *IW3C2 (International World Wide Web Conference Committee)*.

Yannis Bakos, Florenica Marotta-Wurgler, and David R Trossen. 2014. Does anyone read the fine print? consumer attention to standard form contracts. *Journal of Legal Studies*, pages 1–35.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adem Obeng, Stefan Müller, Akitaka Matsuo, and William Lowe. 2018. quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30).

Uri Benoliel and Shmuel I Becher. 2019. The duty to

read the unreadable. *Boston College Law Review*, 60:2255–96.

Tom B Brown. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jonathan H Choi. 2023. How to use large language models for empirical legal research. *Journal of Institutional and Theoretical Economics*, pages 214–43.

Jessica Clement. 2023. Leading android gaming apps worldwide 2023, by downloads. *Statista.com*.

Edward Fichtner. 1980. Measuring syntactic complexity: The quantification of one factor in linguistic difficulty. *Die Unterrichtspraxis (Teaching German)*.

Nelson Francis and Henry Kučera. 1979. Brown corpus manual.

Stefan Gries. 2016. *Quantitative Corpus Linguistics with R*. Taylor Francis.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review.

Nancy Kim and DA Jeremy Telman. 2015. Internet giants as quasi-governmental actors and the limits of contractual conset. *Missouri Law Review*, 80:723–770.

J Peter Kincaid, Robert P Fishburne Jr., Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training*.

Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesa Lagioia, Hans-Wolfgang Micklitz, and Paolo Torroni. 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(32).

Florencia Marotta-Wurgler and Robert Taylor. 2013. Set in stone? change and innovation in consumer standard-form contracts. *New York University Law Review*, 88:240–85.

Eric Martínez, Francis Mollica, and Edward Gibson. 2022. Poor writing, not specialized concepts, drives processing difficulty in legal language. *Cognition*, 224.

Eugène Mollet, Alison Wray, Tess Fitzpatrick, Naomi R Wray, and Margaret J Wright. 2010. Choosing the best tools for comparative analyses of texts. *International Journal of Corpus Linguistics*, 15.

Michael Rustad and Thomas Koenig. 2014. Wolves of the world wide web: Reforming social networks' contracting practices. *Wake Forest Law Review*, 49:1431–1517.

Tim R Samples, Katherine Ireland, and Caroline Kraczon. 2024. Tl;dr: The law and linguistics of social platform terms-of-use. *Berkeley Technology Law Journal*, 39:47–110.

Julia Silge and David Robinson. 2016. tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, 1(3).

Daniel J Solove. 2013. Privacy self-management and the consent dilemma. *Harvard Law Review*, 26:1880–1903.

Isabel Wagner. 2022. Privacy policies across the ages: Content and readability of privacy policies, 1996-2021. *ACM Transactions on Privacy and Security*.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, and 5 others. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Jan Wijffels. 2023. udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the "udpipe" "nlp" toolkit'. *Scientific Data*, 3.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Louiz Bonino da Silva Santos, and et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.

Ethan Zuckerman and Chand Rajenda-Nicolucci. 2021. Deplatforming our way to the alt-tech ecosytem. *Knight First Amendment Institute*.