

CAN LLMs READ PRIVACY POLICIES LIKE LAWYERS?
AN APPROACH TO CORPUS CREATION AND
PERFORMANCE BENCHMARKING

*Florencia Marotta-Wurgler and David Stein**

February 3, 2025

Abstract

Developments in machine learning, particularly the introduction of large language models (LLMs), have generated new opportunities for the automated analysis of legal texts, including contracts. However, the interpretation and classification of legal text can be highly specific and complex, as it requires legal training to learn the law’s internal reasoning, interpretative processes, and unique vocabulary. A central question is therefore how to train LLMs to engage in legal tasks and how to measure their performance.

In this paper, we introduce a labeling and coding approach tailored for training and testing privacy policies using LLMs that capture the context and nuance involved in specific aspects of legal interpretation. We develop and make available a training set of 162 privacy policies, hand-coded over 64 dimensions by legally trained coders, whose coding was supervised and reviewed by one of us, to evaluate the capabilities of off-the-shelf language models “pre-trained” on general corpora of text, such as BERT, as well as those trained or fine-tuned on legal texts, such as LEGAL-BERT. Importantly, our approach incorporates relevant legal rules across the U.S. and the E.U., and it addresses the inherent difficulty associated with interpreting terms that are characterized by inconsistency or ambiguity or are subject to reasonable disagreement in interpretation.

We demonstrate how the dataset thus generated can be used to benchmark the ability of LLMs to interpret real-world privacy policies in a way that captures the process of contractual interpretation. We offer some preliminary results where we “tune” various LLMs to label key aspects of privacy policies and automate our coding process.

* NYU School of Law and Northeastern Computer Science Department and Law School. Thanks to Adam Badawi, Stefan Bechtold, Christoph Engel, Louis Kaplow, Tamar Kricheli-Katz, Michael Livermore, Aileen Nielsen, Steve Shavell, Rebecca Stone, Avishalom Tor, Amit Zac, Eyal Zamir, and participants at the University Bonn Law and Economics Workshop, University of Amsterdam Workshop, CELS 2023, EBEAL workshop, Online Workshop on the Computational Analysis of Law, Cornell Tech DLI Seminar, NYU Law and Economics Workshop, Harvard Law and Economics Workshop, and NYU Privacy Group workshops for helpful suggestions and comments.

I. INTRODUCTION

Recent developments in Natural Language Processing (NLP) have opened promising avenues to the systematic analysis and processing of large amounts of text. Tools using NLP have been used to generate chatbots and virtual assistants. In law, they have been used to increase the efficiency of discovery and document review. Research has explored the extent to which NLPs can handle legal tasks designed for humans, such as taking law school exams (Blair-Stanek et al, 2023), passing the Bar exam (Choi et al, 2021), engaging in statutory reasoning (Blair-Stanek et al, 2023), or identify types of legal reasoning (Thalken et al, 2023). The introduction of large language models (LLMs) and their increased ability to process text has only heightened interest in legal applications, because of LLMs’ ability to process large, unstructured, natural language texts.

Even though these technologies appear to have the potential to conduct more complex legal tasks, such as legal research and interpretation, little is known about their actual ability to do so (Thalken et al, 2023; Guha et al, 2024; Dahl et al, 2024). Interpreting human-written legal texts is challenging due to their nuance, internal logic, specialized vocabulary, dependence on specific contexts, and—not least—potential inconsistencies. LLMs weren’t built specifically for legal use cases, the mechanism that makes them work is only partially understood, and their internal workings are inscrutable. Even qualitative investigations aimed at understanding what LLMs do and “know” are areas of active—and sometimes contentious—research. In sum, an open question is whether LLMs can effectively handle more sophisticated legal tasks; answering this question requires a way to benchmark their performance (Guha et al, 2024; Frankenreiter & Nyarko, 2022).

This paper contributes to the growing literature on creating training datasets for automating the interpretation of legal texts and conducting benchmarking exercises. We introduce a coding protocol and a set of hand-coded privacy policy samples that allow for training and benchmarking LLMs.

Why privacy policies? Privacy policies are an interesting exemplar of the legal use cases to which LLMs may aspire, and they are important in their own right, because they are a crucial feature of any online service or application that handles personal information. They are typically long, complex legal documents that govern the relationship between individuals and firms regarding the use, sharing, and protection of personal information, and the most-used tool to regulate such practices. The U.S.’s “Notice and Choice” regime relies heavily on the disclosures on privacy policies, and much of the European Union’s General Data Protection Regulation (GDPR) requires particular disclosures and explicit consent to firms’ privacy policies. Most recent state privacy regulation, including California’s Consumer Privacy Act (CCPA), also rely on particular disclosures in firms’ privacy policies. They also perform a dual role, as they are often the basis of enforcement actions for violations of the aforementioned laws. But fundamentally, privacy policies are contracts that allocate the rights and risks related to information practices between firms and consumers.

We provide a new method and toolset for labeling the privacy-relevant features of online contracts. Our methodology (which builds on Marotta-Wurgler, 2017) codes privacy policies on 64 dimensions related to the collection, use, sharing, and security of personal information, dispute resolution, control, and contract modification terms that relevant regulatory regimes, including FTC guidelines and GDPR. For example, several questions focus on what type of information is collected and the purposes of collection. This methodology allows us to collect

additional information that is useful for working within the limitations of the current generation of commercially available LLM-based systems, such as supplying relevant legal rules that are relevant in interpreting the meaning of terms.

We then use this methodology to code 162 privacy policies, terms of use, and any other document incorporated by reference in the privacy policy (such as cookie policy or CCPA compliance link). The sample firms are spread across seven different markets where information sharing is salient to consumers: cloud computing, social networks, dating sites, gaming, news and reviews, special interest message boards, and adult sites, as well as from each tier-1 industry category in the IAB Content Taxonomy 2.0 (IAB Tech Lab, 2024).

For each policy, coders with legal training highlighted any text relevant to one of the 64 “questions” or dimensions of the privacy policies. (A simple example: Does the privacy policy or terms of use have a class action waiver?) This enables us to capture context in an efficient manner and also gives us a glimpse of how legal texts are read. Coders then answer the question based on a comprehensive set of choices, including the possibility that the policy is silent on the question (as silence acquires different meanings depending on whether there are default rules governing that particular exchange). Coders are also asked how confident they are in their answers. Each policy was independently tracked by two rising 2L students, and the answers were reviewed by one of us during weekly meetings during a 10-week period, where we discussed sources of disagreement and difficulty in interpreting the questions or the text. We revised the wording of some questions, or changed the answer set to reduce disagreements that may have resulted from the coding process itself. Naturally, disagreements remained. Like most legal documents, privacy policies are sometimes internally inconsistent or include terms that are ambiguous or lend themselves to reasonable differences in interpretation. We treat these sources of disagreement as a feature of the process of legal interpretation, rather than a bug. This generated a rich training dataset in which each policy was coded and highlighted twice. The hand-coded dataset reveals interesting differences in processing legal documents across coders, underscoring the importance of context in ascertaining the legal meaning of terms.

Next, we use the dataset to benchmark existing LLM technologies’ ability to perform specific legal interpretation tasks, including identifying “difficult” questions. We offer preliminary results by running our coding through various off-the-shelf NLPs: GPT-4 and Claude 3. We also evaluate NLPs’ abilities to replicate the highlighting of relevant text using LEGAL-BERT and BERT-BASE. Like the research that shows that such models experience large degrees of legal errors, or hallucinations (Dahl et al, 2024), we find that untrained NLPs perform well in simple questions, but poorly at interpreting more complex text in privacy policies.

II. BACKGROUND AND PRIOR EMPIRICAL RESEARCH

Recent developments in NLPs offer promising new opportunities to automate the analysis of legal texts such as contracts. Past efforts have focused on analyzing the content of privacy policies, as these documents have been at the forefront of regulatory efforts such as the European Union’s General Data Protection Regulation (GDPR) and govern the collection, use, sharing, and security of personal information between firms and consumers.

Empirical research on privacy policies has relied on both hand-coded and automated data sets. The studies that rely on hand-coded data sets document the content of privacy policies and the role of market forces in shaping terms (Marotta-Wurgler (2017), how privacy policies change over time (Davis & Marotta-Wurgler,

2020), and how GDPR may have affected US-facing information practices (Frankenreiter, 2022; Davis & Marotta-Wurgler, 2024). Other studies have explored the terms in privacy policies to document how they became longer over time (Amos et al, 2021). Studies using some automated approaches have explored the extent to which policies comply with GDPR (Lippi et al 2019), Kubicek et al, 2022), (Peukert, Bechtold, Batikas, 2021) (Becher & Benoliel, 2021), Frankenreiter (2022).

Recent work on privacy policies has focused on automating their analysis using NLPs. To this date, there have been several initiatives on this front, which fall broadly into four main categories:

1. *Creating Datasets*

One set of research focuses on the collection of documents. Collecting online contracts from across the internet is a research task in its own right. No legal or technical standard dictates where a company should display its policies and licenses, how they should be formatted, whether they should be scattered across multiple pages, or if they can change dynamically based on a user’s location or identity. Capturing online contracts can also involve checking for updates over time, or drawing from archival services like the Internet Archive when records are available. The largest collection of privacy policies, the Princeton-Leuven Longitudinal Corpus of Privacy Policies, extracted text from 130,000 websites’ privacy policies, using selection samples from a twenty-year period. This is the largest dataset on privacy policies, but it is limited to the “front page” of the privacy policy, meaning any text incorporated by reference is not included.

2. *Building Training Datasets*

Machine learning (ML) requires training data, which usually requires some amount of manual human labeling. ML techniques use statistical methods to emulate some modeled behavior. For reading legal documents, that means compiling examples of correctly-labeled text. The design and construction of training data often sets the direction of ML research. The paper that set of the current flurry of NLP research also was working towards an independently-defined benchmark. Creating training data and benchmarks can push ML research in certain directions, and the absence of training data for a given task makes ML development difficult.

The most notable training data in privacy policies and online contracts is maintained by the Usable Privacy Project,¹ which maintains several privacy-related corpora (Wilson, 2016). The OPP-115 dataset includes 115 website privacy policies with detailed, word-level annotations. The labels were created by showing trained student labelers individual paragraphs and asking them to identify phrases within that paragraph that fall into one of ten categories, making each phrase with a category, subcategory, and annotation. For example, one coder was shown a paragraph containing the phrase “... NOTMC and its agents may collect some information that identifies you ...” and marked the phrase “may collect” as “<category:first party collection> <subcategory: does or does not> <annotation:does>.” The OPP-115 data set has been used for benchmarking to evaluate the extent to which LLMs can engage in legal reasoning (Guha et al, 2023).

New machine learning techniques and applications are developed using that training data. Since NLP is a relatively young field and the ability to read online contracts unlocks a relatively narrow range of potential applications, online contracts

¹ See Usable Privacy Project, available at <https://usableprivacy.org/>.

have not been a central feature in NLP research. Several projects have made progress in accurately predicting which annotations the OPP-115 labelers selected. Notably, the Polisis system used the OPP labels to annotate the Princeton-Leuven Corpus (Mousabi et al 2020). Some EU-based projects have trained ML systems on OPP-115, using OPP labels to pick out GDPR-relevant clauses with reasonable success. Other projects, such work by CLAUDETTE (Lippi et al, 2019), discussed below, developed their own training data to answer targeted questions.

Finally, there is a dataset related to privacy policies that labels the content of cookie banner disclosures related to stated purposes for data collection by focusing on longer legal text (Santos et al., 2021).

3. *Applied ML*

Other research projects leverage the Polisis software to investigate the content of privacy policies (Harkous et al, 2018; Okoyomon et al, 2019), the inherent risks in information practices, and how privacy policies are affected by regulations such as GDPR (Nejad et al, 2020; Linden et al, 2019; Zaeem et al, 2021). Not all projects employ ML techniques.

In the privacy policy space, the OPP-115 dataset has been used create to tools for extracting specific clauses from privacy policies (Mousavi Nejad et al., 2020) and to generate related datasets, either by adapting its annotations for new tasks like question-answering or GDPR compliance (Poplavska et al., 2020; Ahmad et al., 2020), or as an input into composite legal-task benchmarks such as LEGALBENCH and PRIVACYGLUE (Guha et al., 2023; Chalkidis et al., 2022). The OPP taxonomy scheme has further served to organize other privacy-related datasets (Ravichander et al., 2019).

More generally, there is a growing area of research that uses ML techniques to analyze terms in contracts beyond privacy policies. These include Rauterberg & Talley (2017); Talley & O’Kane (2012) (investigating force majeure provisions in merger agreements); Nyarko (2019) (using supervised machine learning to investigate the frequency of choice-of-law provisions in certain agreements); Alschner (2017) (using a procedure identifying keywords to examine the impact of investment arbitration on investment protection treaties). Kosnik (2014) develops different measures of the “completeness” of contracts to analyze differences in flexibility of agreements. Beuve et al (2019) and Moszoro et al (2016) investigate the differences in rigidity of private and public contracts). McLane (2019) examines the costs and benefits of using boilerplate language in SEC disclosures using various measures of boilerplate.

III. THE ROLE OF DATASETS AND BENCHMARKS IN AI DEVELOPMENT

A. *A Quick Introduction*

Machine learning (ML) is a subfield of artificial intelligence focused on developing statistical algorithms that enable computers to learn from data without being specifically programmed to do so. Rather than writing instructions for a computer program to follow, ML engineers “train” their programs using examples of the kind of output they’d like the program to produce for a given input. Artificial intelligence (AI) refers to any system that automates a behavior that requires

intelligence, but the term is often used to refer specifically to computer programs created using ML techniques.

Rather than programming algorithms directly, machine learning researchers design AIs to emulate behavior modeled in datasets. So, while there are technical limitations on the *kind* of behavior an AI can emulate, the specific behavior of an AI comes from the datasets used to build, modify, and test that AI.² To make this explicit: two identical AIs trained on datasets modeling different behavior would behave differently. Two different AIs trained on datasets modeling identical behavior would similarly.

B. *The Importance of Benchmarks*

Training, modification, and tuning of AIs requires two things: a set of examples (the dataset) and some way to measure the difference between an AI’s output and the example output (a metric). AI benchmarks are typically defined using one or more datasets and metrics.

Selecting the appropriate metric and dataset is highly context specific. For example, many facial recognition systems work by computing the “distance” between pictures of faces. An appropriate dataset might contain pictures of the same group of people in a variety of different angles and lighting conditions, and a simple but reasonable measure might subtract the maximum distance between pictures of the same person from the minimum distance between pictures of different people, which would give a higher score when the AI clusters pictures of the same person together. NLP tasks use a wide variety of metrics depending on the format of the dataset, the kind of task the AI is being trained to perform, and the kind of technologies the AI uses to learn.

AI benchmarks from domain-specific tasks require domain-specific datasets, which necessarily require some level of participation from subject-matter-experts. Even if an AI can be trained on more generic data, validating that AI’s capabilities requires some established benchmark. In the legal context, a major challenge in the automatic interpretation of contracts is creating a structured representation of the meaning of terms against pertinent legal rules, and how the terms relate to one another, which Frankenreiter & Nyarko (2022) call “legal ontologies.” Many LLMs are trained on datasets that include contracts and judicial decisions and may be capable of capturing these legal ontologies. To this date, few datasets provide example input/output pairings that capture those features, meaning we have no meaningful ground truth from which any metric could measure an AI’s ability to capture those features of legal texts. Recent efforts have begun to explore the ability of LLMs in performing legal tasks.

Building systems capable of automating legal reasoning and assessing their ability to perform legal tasks in practice requires access to benchmarks that accurately represent those tasks. Recent interdisciplinary efforts by computer scientists and legal academics have begun compiling existing legal reasoning benchmarks into a common framework. Guha et al (2023) construct a legal reasoning benchmark to evaluate LLMs ability to engage in six different types of legal reasoning. Zheng et al (2021) use hand-coded data sets to determine those instances when pretraining LLMs may result in better performance of legal tasks.

However, these underlying benchmarks are built around hand-coded data sets designed for more constrained technologies, such as data sets with clause-level annotations, stylized vignettes, and simplified legal questions that may not capture

² [CITE]

most of the nuance or complexity that arises in legal practice. In a contractual setting, for example, this often involves identifying collections of interdependent clauses scattered across a collection of documents, and ascertaining their meaning on one of more targeted legal questions. To assess the capability of AIs performing general legal interpretation, we need to compile new datasets tuned to the kind of targeted questions and holistic reading and interpretation paradigmatic of legal practice.

1. *Tuning*

Training an AI from scratch, especially an AI like an LLM, can require an almost inconceivable amount of training data and computation. By contrast, it is possible to create a high-performing AI with a relatively small amount of data by updating an already-trained AI or repurposing parts of an already-trained AI when creating a new one. This process, sometimes called *tuning*, can repurpose existing AIs for similar tasks that have different data or outputs. In our dog-breed-identifier example, the pre-trained AI might already have subcomponents trained to filter the background of an image or distinguish between eyes, paws, and fur. Tuning that AI to recognize cats might require orders of magnitude fewer examples than starting from scratch.

As with testing, tuning requires a well-defined benchmark. AIs are trained and tuned by iteratively adjusting the AI to generate “better” outputs. AIs are more likely to achieve high performance if the dataset and metric used for tuning can accurately distinguish between output that is horribly wrong versus merely mediocre.

Given modern LLMs’ impressive ability to process English text, the prospect of tuning LLMs for specific legal tasks seems particularly promising. Even if LLMs can’t provide high-quality legal automation out of the box, a well-defined dataset and properly selected metric could facilitate the creation of trustworthy LLMs tailored for specific legal tasks.

IV. DATA AND METHODOLOGY

We contribute to this literature by offering a tool kit and dataset designed to capture context and nuances in interpretation of privacy-relevant provisions in privacy policies. Our sample includes the policies, terms of use, and documents incorporated by reference therein, including GDPR statements, cookie policies, and CCPA disclosures, of 162 firms (a combination of a subsample of 261 firms analyzed in Marotta-Wurgler (2017) as well as a randomized sub-sample from the firms in OPP-115), from seven online markets where consumers often share personal or sensitive information: adult, cloud computing, dating, gaming, news and reviews, social networks, and special interest message boards, as well as from each tier-1 industry category in the IAB Content Taxonomy 2.0. These are markets where information sharing is relatively more salient than in others where information sharing is a secondary aspect of the particular transaction, such consumer retailers or news sites. There are interesting differences in privacy concerns across these markets.³ The firms involved do business in the United States but almost always also have overseas operations. They include giants like Facebook and Google, and many smaller firms like veggiedate.com.

³ Our sample is smaller because of resource constraints: our improved coding scheme involves dozens of additional terms and contracts have only become longer and more complex since 2017.

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

We track 64 terms across nine different information privacy categories and 33 sub-categories. These include terms related to Notice (13 terms), Sharing (7), Security (8), User Control (8), Enforcement (8), Privacy by Design (2), Data Practices (1), CCPA related terms (10), GDPR-only related terms (5) and contract related terms (1). The questions and answers are described in Table 1 in the Appendix. The terms span from tracking contract formation rituals and compliance with the notice and other requirements of GDPR, to terms related to dispute resolution, such as whether the policy includes a choice of law, forum, or class action waiver.

A major challenge in the automatic interpretation of contracts is creating a structured representation of the meaning of terms in context and against the relevant legal rules. Many of the provisions in our coding often require reading clauses that incorporate other documents and clauses by explicit or implicit reference, so we capture them all.

Questions	64	
Categories	11	
Total Coders	18	
Coders per Policy	2+	1
Policies	88	74
Paragraphs	29,359	27,372
Words	937,943	977,364
Highlight Annotations	14,811	7,218
Policy Classifications	11,718	4,733
Confidence Scores	9,608	4,464

Table 1

A. Labeling Methodology

Each privacy policy and documents incorporated by reference, including Terms of Use, Cookie Policy, CCPA disclosures, and GDPR addenda was coded by a pair of rising 2Ls who had been trained to read and code privacy policies and understood the legal rules that govern them. A total of 16 students coded the contracts. The two of us reviewed the contracts. Table 1 reports summary statistics. We coded 64 questions spanning 11 categories of information privacy practices and other terms defining the relationship between data processors and users, resulting in a total of 162 coded policies, Terms of Use, and related documents for a total of 56,731 paragraphs.

Our coding process involved answering questions that listed a comprehensive set of answers in a multiple-choice format. For each question, coders are asked to highlight any text they consider relevant in answering the question, select the answer, and report their confidence in their answer. Each document was coded by two J.D. students, and their answers were subsequently reviewed by one of us. The final coding is the result of an iterative process designed to reduce disagreements between

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

coders resulting from difficulty interpreting the questions or matching the policy text to the available answer sets.

Figure 1 presents a snapshot of the coding tool, where coders can scroll and highlight the privacy policy and related documents set of contracts on the left panel while answering the questions on the right panel.

The screenshot shows a coding tool interface. On the left, a document titled "DOCUMENT B - Facebook" is displayed with several paragraphs of text. Paragraph 21 is highlighted in green, and paragraph 25 is highlighted in yellow. On the right, a question titled "GDPR" is shown. The question asks: "63. GDPR-Automated Processes (682_2020) 14 sentences confidence: low". Below the question, there is a table of options to select from:

please select the most appropriate option		
<input type="checkbox"/> A.7(4)	<input type="checkbox"/> A.31(2,3)	<input type="checkbox"/> A.32(2)
<input type="checkbox"/> A.34(4)	<input type="checkbox"/> A.35(2)	<input type="checkbox"/> B.21(1,2,4)
<input type="checkbox"/> B.23(1,3)	<input type="checkbox"/> B.32(2)	<input type="checkbox"/> M.29(1,4)

Below the table, there are radio buttons for "no", "yes", "not applicable", "does not disclose", and "other (enter here)". At the bottom, there is a "confidence" section with buttons for "very low", "low", "medium", "high", and "very high", and an "additional comments" field.

Figure 1

Specially, the process of coding is conducted through a tool that simultaneously loads the policy as well as the coding taxonomy (the 64 term variables). When measuring each policy, coders must answer each taxonomy question by selecting one of several answers. For this, coders are required to read the whole document in making a particular coding decision. As a simple example, one question in the taxonomy is whether the privacy policy includes a change of terms clause. Coders must read the loaded contract and find the relevant part of the policy that addresses contract modification and highlight the parts of the text that form the basis of their answer. Instead of labeling, the highlighted text is linked to the question being addressed. This allows us to link the question to the relevant text in the policy.

Once the relevant text is highlighted, coders must record the answer indicating whether the contract includes a modification clause, states that no modifications will be allowed, or is silent on the issue. The exercise of reading the entire document and highlighting the relevant parts to code each specific term allows coders to take context into account—both within the contract itself by not forcing answers from specific parts of the document but rather from a more comprehensive reading, and by coding terms against meaningful benchmarks. This makes the exercise of identifying relevant language and translating it to a quantifiable format “holistic,” the approach of a trained lawyer.

An advantage of our coding methodology over plain labeling is that it allows for coding “outside the contract.” For example, our coding method allows us to note when a contract is silent on an issue—a frequent occurrence—and, unlike existing approaches, interpret the meaning of this silence. Silence has different meanings in different circumstances, and coding methodologies that look only within the legal document to perform legal analysis are unable to account for that. In some contexts, silence implies that the relevant default rules allocate rights and risks among the parties; in others, silence may evidence a violation of a rule for failure to disclose a mandatory term. Measuring policies against relevant background rules performs three important functions. First, it provides a template to measure the extent to which firms adhere to guidelines and mandatory terms. Second, it

allows for meaningful comparisons across firms and markets. Third, it allows us to ascertain silence in a legally meaningful way.

The coding process itself is also designed to capture some of the nuances of legal reasoning. We do so by teasing out differences in coding and highlighting of the same document. For each policy, we tracked the amount of time it took each coder to complete the questions, the text each coder highlighted as relevant when answering each question, the answers, and their self-reported confidence using a Likert scale. This generated a rich data set in which each policy was coded and highlighted twice. The result was 162 fully coded policies and related documents on 64 dimensions, comprised of 56,729 paragraphs, of which 22,029 were highlighted by 18 coders, including ourselves.

In theory, coders should agree when answering each question. Let's return to our example question asking whether the privacy policy has a change of terms clause. Ideally, coders annotating the same policy should read it and highlight the same relevant text. In our example, this would be a clause that states something like: "We may update this Privacy Policy from time to time. When we update the Privacy Policy, we will notify you by updating the 'Last Updated' date at the top of the new Privacy Policy, posting the new Privacy Policy, or providing any other notice required by applicable law. We recommend that you review the Privacy Policy each time you visit the Platform to stay informed of our privacy practices." Next, coders should select an answer that indicates that the policy indeed contains a modification clause and should also note how certain they are of their response. This is an easy question, so correspondence is likely. Other questions, such as whether the data are shared with third parties or whether the individual can control what is collected and/or shared, may not be as straightforward. Privacy policies often include terms that are inconsistent or include rights that are subject to modifications, making it hard to understand what the underlying rights and obligations are. We expect some inconsistencies among coders in such situations.

B. *Iterative Revisions*

In addition to needing to account for background rules and interdependent clauses, automated analysis of contracts is further complicated by the ambiguity that sometimes exists in legal documents. In some cases, this ambiguity may be resolved with applicable rules of interpretation. In others, it cannot. Ambiguity can also result from the coding schema itself. Ambiguous or vague labeling can yield indeterminate answers. Finally, some questions are just hard to tackle systematically. Even sophisticated readers may reasonably disagree. Except for ambiguities created by the coding schema, these problems are inherent in the exercise of legal reasoning. Our coding implementation design, based on iterative revisions, seeks to remove sources of coder disagreement resulting from ambiguity in our questions while capturing those disagreements resulting from the process of legal interpretation.

During an iterative period when we met with each pair of coders once weekly over a ten-week period, we tracked all answers and identified the instances where coders were likely to disagree in their responses and identified the causes for such disagreements. We found that sources of confusion or disagreement from the coding protocol itself stemmed from: (1) coders interpreting the question in conflicting ways due to confusing wording, or (2) the answer choice set for a given question did not fully map onto the text or law, or properly account for the existence of default rules that may alter the meaning of contractual silence. When appropriate, we reworded the questions to make them clearer, or made the answers more granular to remove any ambiguity that may have been inherent in the questions themselves.

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

We repeated the exercise until we no longer saw a reduction in inconsistencies among coders. Yet disagreements remained. Some were the result of coder mistakes, which we corrected during revisions. Most other, however, were the result of: (1) difficulty in interpreting inconsistent or ambiguous clauses, or (2) reasonable disagreements about hard questions. We exploit differences in highlighting and answer choices for the same question to identify ambiguity and the presence of “hard questions.”

Our approach is different in this regard. Almost every attempt at manual text annotation and classification has made an effort to minimize inter-coder disagreement (Wilson, 2016). Extant privacy policy datasets discard points of inter-coder disagreement during construction. But indeterminate interpretations or differences in conclusions resulting from legal analysis of contracts (or other legal texts) are an inherent feature of this process. Thus, cases where expert coders disagree or feel unsure should be expected and preserved. Ideally, a well-trained LLM would report similar uncertainty. Our method aims to ensure that disagreements and uncertainty in the manual coding reflects ambiguity in the text and not the schema or instructions. Our methodology thus hopes to offer a novel application of “agile” techniques to reduce sources of ambiguity exogenous to the sample text.

C. Insights Into the Process of Legal Analysis

The process of creating a corpus using our granular approach revealed some interesting findings. First, even after repeated iterations to reduce ambiguity by revising the wording of the questions, coder agreement varied at the term and term category level, suggesting that some disagreements were the result of ambiguities in the text or reasonable disagreements about the meaning of the terms.

Indeed, we found a positive correlation between coder agreement and self-reported confidence on each question evidenced in Figure 2, which plots Cohen’s Kappa against average self-reported confidence. We hypothesize that coders are aware when they face hard questions instead of being misguided by poorly drafted questions.

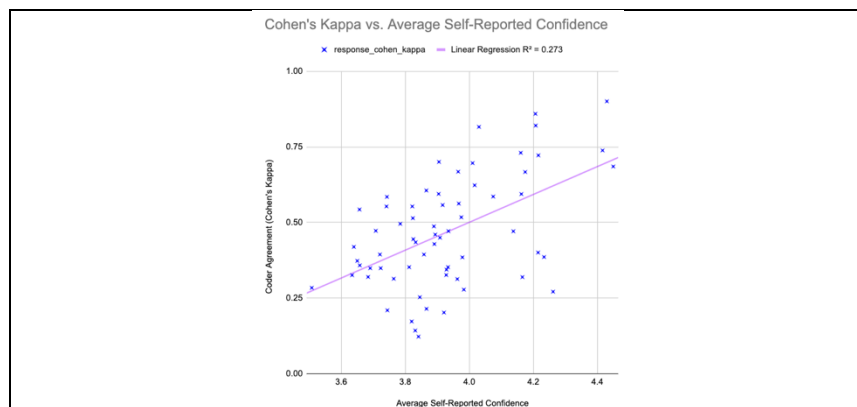


Figure 2

We also found that coder correspondence varied for sets of terms, revealing that some terms were easier to ascertain than others. This is evidenced in Figure 3. Terms related to dispute resolution, including terms tracking whether the policy included choice of law, forum, and class action waivers, had naturally high degrees of correspondence among coders, since the presence or absence of these terms is straightforward. In contrast, terms related to data practices, which can be internally inconsistent or complex, resulted in more mismatches.

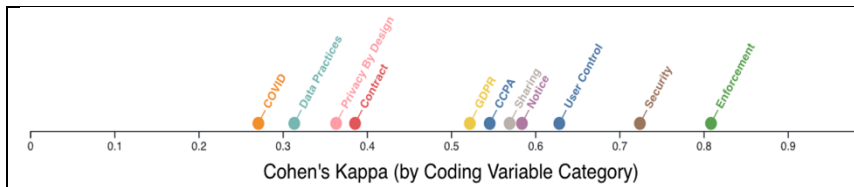


Figure 3

Finally, we found that coders expressed variations in their highlighting approaches, highlighting multiple terms under some circumstances to answer particular questions, suggesting that they read the contract “holistically” but also revealing through their analysis that the terms in privacy policies, and in most contracts, constitute an interconnected eco-system and that meaning can be best ascertained when reading the contract as a whole.

V. CASE STUDIES: PROOF-OF-CONCEPT PRACTICAL APPLICATIONS

Using our dataset, we evaluate models on their ability to perform two tasks. The first task examines the extent to which an off-the-rack LLM can perform the task of evaluating the terms in privacy policies and coding the terms using our methodology, as listed in Table 1 in the Appendix by selecting the most likely answers. Second, we evaluate highlight prediction: given a targeted legal question and a paragraph from a relevant legal document, we evaluate the LLMs ability to predict whether a legal annotator would mark that paragraph as containing information relevant to answering the question. For each task, we measure our results across the entire dataset. We also present results by category and by question.

Table 2 reports the result of the coding approach by GPT-4 and Claude 3. The first task, which we label “holistic classification,” is a multi-classification task that uses the entire policy as input: given our questions and a policy from our dataset, select the most likely answers for each question.

We evaluate the holistic classification task using batched cross-entropy loss. Each of the k policies in the dataset is associated with n sets of labels, $\{L_1, L_2, \dots, L_n\}$, each corresponding to a question in our coding scheme. For each label L_i , has set of options M_i . Given the ambiguity present in some privacy policies, the ground-truth value of L_i^k may not be a single value, but rather a probability distribution over M_i . Coder responses are therefore definitionally noisy. We compute the goal probability distribution y_i^k as $(c^{k_{i1}}, \dots, c^{k_{im}})$ normalized to sum to 1, where $c^{k_{ij}}$ represents the number of coders who selected option j for question i on document k . We apply label smoothing to account for noise, as described in (Müller et al.,

2020), setting α to .1. We use LLMs to generate a probability distributions p^k_i over the set of options for each label. When logprobs are available, we generate the distribution by crawling the response tree of each branch until an answer is selected or the net probability is negligible. We evaluate model responses by computing binary cross-entropy loss between model response and the reference distribution. We record the average loss by question, category, and across the entire dataset.

Because some policies contain more than 32 thousand words, we can only test LLMs with sufficiently large context windows without resorting to context-expanding techniques or alternative models, which are out of scope for this project. Table 2 reports the performance of the major commercial LLMs with sufficiently large context windows. We explore the performance of Claude 3 and GPT-4 and compare the outcomes to random guessing using zero-shot labelling. We find that the models perform relatively well in straight-forward, discrete questions, such as whether the policy states that the data collector tracks data for COVID-related data, or whether it includes certain dispute resolution clauses, such as class actions. However, performance becomes closer to random (albeit better) as the questions become more complex and rely on interpreting various provisions that modify each other, such as terms related to information sharing and data security.

Category	Average BCE Loss		
	Claude 3 Haiku	GPT-4	random guesses
overall	0.292	0.212	0.467
CCPA	0.301	0.202	0.475
COVID	0.007	0.007	0.476
DP	0.098	0.333	0.503
E	0.180	0.171	0.474
GDPR	0.253	0.168	0.441
K	0.112	0.086	0.413
N	0.360	0.202	0.456
PBD	0.254	0.302	0.473
SE	0.254	0.245	0.488
SH	0.333	0.296	0.463
UC	0.406	0.204	0.467

Table 2: Average cross-entropy on holistic classification task, by category. At time of writing, few publicly available models have a large enough context window (over 32k tokens) to perform the task.

Using our dataset, we evaluate *highlight predictions*, a binary classification task, by concatenating individual paragraphs with question text and option descriptions. $y_{ij} = 1$ if at least one coder flagged paragraph j as relevant when answering question i , and 0 otherwise. We tested zero-shot labeling, prompting the model to answer whether the paragraph was relevant and computing the relative likelihood of an

affirmative or negative response using the tree-crawling approach described above. The zero-shot accuracy, precision, and f1 scores of several models are shown in table 3.

We note that performance varies significantly across categories and questions, including questions within the same category. While differences in performance between models may be an artifact of our prompt design, we found the variance between similar questions about similar topics striking. At least for the systems we tested, an LLM’s ability to answer one legal question appears to not be predictive of that LLM’s ability to answer other questions, even within extremely narrow domains like “properties of sharing practices” described within a privacy policy.

model	acc.	recall	f_1
LEGAL-BERT	87.16	14.04	1.58
BERT-BASE	63.69	22.05	1.73

Table 3: Average zero-shot performance on highlighting task, optimizing for f1. Because highlighting is noisy and heavily skewed, we suspect a certain number of false positives are unavoidable.

VI. DISCUSSION

There are several proposals to use LLMs for legal interpretation and drafting tasks. Our results show that, while LLMs are capable of answering specific questions with a reasonable degree of accuracy, that accuracy is brittle, and may not generalize to tasks that seem facially analogous. LLMs can still be useful and can still be tuned and validated to perform a wide range of narrow, well-defined tasks that would have been unimaginable for computers a few years ago.

Our goal is to contribute to ongoing efforts to perfect LLMs in the process of performing legal tasks, such as contract interpretation. We offer a hand-coded corpus that tracks some of the nuances of legal interpretation by coding terms in an organic way as opposed to by rigid labeling of individual paragraphs. Our coding approach seeks to bring context into account by including all relevant legal documents incorporated by reference in each privacy policy and by bringing in the relevant law in the construction of questions and answer set.

This project is designed to contribute to the growing body of legal task corpora, and will be added to the LegalBench consolidated corpus.

CONCLUSION

The automation of legal tasks is attractive to both legal and technical practitioners. In the legal domain, automation might improve the quality and efficiency of legal practice, it might also help to address the United States’ struggles

in providing access to justice. For NLP researchers, the law's complex, functional use of natural language to create documents with observable effect presents a challenging interpretation task that produces results that are susceptible to objectively defined validity measurements.

Building systems capable of automating legal reasoning and assessing their ability to perform legal tasks in practice requires access to benchmarks that accurately represent those tasks. Recent scholarship has begun compiling existing legal reasoning benchmarks into a common framework. However, these underlying benchmarks are built using techniques designed for older technologies that do not capture the complexity of some legal tasks. We offer a toolkit and hand-coded corpus of privacy policies that leverages some of the increased capabilities of more current NLPs by providing more granular coding and replicating the task of contractual interpretation in ways that more closely resemble the actual practices. First, it incorporates relevant documents by reference as well as relevant background rules. Second, it allows coders to read the agreements organically without having to label or annotate segmented paragraphs. Finally, the questions themselves have been edited to remove any misunderstanding. We find that coders read contracts holistically when tasked with answering specific questions.

To assess the capability of AIs performing general legal interpretation, we need to compile new datasets tuned to the kind of targeted questions and holistic reading paradigmatic of legal practice. This paper provides a legal dataset for analyzing the content of online privacy policies and terms of service. In constructing that dataset, we demonstrate a method for constructing a coding scheme that maps to practical legal questions, and a coding process that accounts for legal nuance absent from existing dataset. We select privacy policies because they are publicly accessible, map to well-defined legal questions, and are already the target of automated legal interpretation in both research and practice.

APPENDIX: TABLE 1.

Question ID	Category	Subcategory	Question
CCPA-1	CCPA	Notice	Does the privacy policy include a link to the CCPA section, as opposed to in the same privacy policy?
CCPA-10	CCPA	Opt Out	Does the privacy policy offer consumers or users the right to opt-out of selling personal information to third parties with a visible, direct link to "Do Not Sell My Personal Information"?
CCPA-2	CCPA	Notice	Does the privacy policy state that the firm's CCPA policy only applies to California residents? For example, does it include a statement similar to the following one: "This California section supplements the Privacy Policy and applies solely to California consumers (excluding our personnel). The Table below describes how we process California consumers' personal information (excluding our personnel), based on definitions laid out in the California Consumer Privacy Act ("CCPA")."
CCPA-3	CCPA	Notice	Does the privacy policy include a California Privacy Rights Section that explains all rights afforded to users and consumers under the CCPA? For example, these include: the right to request disclosure of business' data collection and sales practices , the categories of personal information collected, the source of the information, use of the information and, if the information was disclosed or sold to third parties, the categories of personal information disclosed or sold to third parties and the categories of third parties to whom such information was disclosed or sold; The right to request a copy of the specific personal information collected about them during the 12 months before their request (together with right 1, a "personal information request"); The right to have such information deleted (with exceptions); he right to request that their personal information not be sold to third parties, if applicable; and The right not to be discriminated against because they exercised any of the new rights.]
CCPA-4	CCPA	Notice	Does the privacy policy directs California Residents to the CCPA section when describing general, non-California exclusive, data practices?
CCPA-5	CCPA	Notice	Does the privacy policy offer California residents an opportunity to request all information shared with

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

Question ID	Category	Subcategory	Question
			third parties in the last year?
CCPA-6	CCPA	Notice	Does the privacy policy offer California residents a direct link via which to contact site and request information?
CCPA-7	CCPA	Notice	Does the privacy policy offer data requests by consumers or users explicitly free of charge?
CCPA-8	CCPA	Data Sharing	Does the privacy policy list the categories of personal information sold in the past 12 months?
CCPA-9	CCPA	Data Choices	Does the privacy policy identify at least two methods for submitting a personally identifiable information or erasure request, in accordance with CCPA? These must include, at a minimum, a web page and a toll-free telephone number.
COVID-1	COVID	(none)	Does the privacy policy include any terms related to contact tracing, health tracking, or other terms in relationship to COVID?
DP-1	Data Practices (DP)	Retention	Does company have a procedure for safely disposing unused or no longer needed data or personally identifiable information?
E-1	Enforcement (E)	(none)	Does the privacy policy provides means by which user can contact the company with any privacy concerns or complaints? Please select all that apply.
E-2	Enforcement (E)	Dispute Resolution	Does the privacy policy or the Terms of Use have a forum selection clause? If so, which forum?
E-3	Enforcement (E)	Dispute Resolution	Do the privacy policy or Terms of Use have choice of law clause? If so, which law?
E-4	Enforcement (E)	Dispute Resolution	Do the privacy policy or Terms of Use have an arbitration clause?
E-5	Enforcement (E)	Dispute Resolution	Do the privacy policy or terms of use have a class action waiver?
E-6	Enforcement (E)	Liability	Do the privacy policy or terms of use disclaim liability for failure of security measures?
E-7	Enforcement (E)	Oversight	Does the privacy policy provides a link to the Federal Trade Commision's Consumer Complaint Form or does it include t he FTC telephone number?

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

Question ID	Category	Subcategory	Question
E-8	Enforcement (E)	Oversight	Does the privacy policy include a privacy seal, certification, or industry oversight organization, other than those mandated by international law, such as the Swiss Privacy Law? Privacy Seals are independent, third-party enforcement programs to monitor company practices and enforce privacy policies. They are designed to provide protection to consumers by allowing Web companies to standardize privacy policies. Privacy seal programs include, among others, TRUSTe, BBBOnline, and CPA Webtrust. These are different from regulatory compliance seals, such as those that the company complies with COPPA, the Children Online Privacy Protection Act).
GDPR-1	GDPR	Notice	Does the privacy policy states that it complies with GDPR or it includes section on GDPR compliance?
GDPR-2	GDPR	Notice	Does the privacy policy state that it complies with EU-US Privacy Shield?
GDPR-3	GDPR	Notice	Does the privacy policy state that GDPR terms apply only and exclusively to EU residents?
GDPR-4	GDPR	Automated Processes	Are users or consumers able to object to the processing or automated decision making that could impact them? This is only applicable if company does profiling or any other automated decision making, such as algorithmic decision making, or any automated decisions that don't involve a human.
GDPR-5	GDPR	Automated Processes	If the privacy policy state that the firm engages in automated decision making, does it provide meaningful information about the logic involved, or significance or effect of such decisions?
K-1	Contract (K)	Contract	Do the Terms of Use or Terms of Service incorporate the Privacy Policy by reference?
N-1	Notice (N)	Type of Data Collected	Does the Privacy Policy include the company's cookie policy, such as an explanation on the text, or a hyperlink to a document with a cookie policy?
N-10	Notice (N)	Change of Terms	Does the privacy policy require the user/consumer to explicitly assent to any material changes?
N-11	Notice (N)	Change of Terms	Does the privacy policy states that material changes made to the policy will be retroactive or apply to previous data collection?

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

Question ID	Category	Subcategory	Question
N-12	Notice (N)	Access	Does the privacy policy summarize the key terms at the top of the policy? Just a table of contents doesn't count.
N-13	Notice (N)	Use of Data	Does the privacy policy explain any data procedures if company is sold or otherwise ceases to exist by, for example, filing for bankruptcy?
N-2	Notice (N)	Type of Data Collected	Does the Privacy Policy note or explain that the company uses tracking elements other than cookies, such as local storage cookies, browser fingerprints, or other non-cookie tracking elements?
N-3	Notice (N)	Type of Data Collected	Does the privacy policy state that the company collects or stores biometric information, such as facial scans, fingerprints, facial patterns, voice or typing cadence?
N-4	Notice (N)	Use of Data	Does the Privacy Policy include a statement noting that that personally identifiable information will be used internally only for business purposes, such as for effecting, administering, or enforcing a transaction, or for sending future correspondence to the user, or for research, internal database compilation, or for servicing the website? Not that using the data for advertising is not considered an internal business purpose.
N-5	Notice (N)	Use of Data	Does the privacy policy include a commitment by the company to use personally identifiable information only for stated, context specific, purposes? These are purposes that a user would expect in the context of the service provided, such as users expecting their personal profiles made available to other users in a dating site?
N-6	Notice (N)	3rd parties	Are third parties allowed to place advertisements that may track user behavior?
N-7	Notice (N)	3rd parties	Does the privacy policy identify third party recipients of shared or sold data?
N-8	Notice (N)	3rd parties	Does the privacy policy define words such as "affiliates" or "third parties," if it uses them?
N-9	Notice (N)	Change of Terms	Does the privacy policy include a "Change of Terms " or modification provision that allows the firm to change the privacy policy?
PBD-1	Privacy By	Oversight	Does the privacy policy require periodic compliance

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

Question ID	Category	Subcategory	Question
	Design (PBD)		review of structural and technological data security measures?
PBD-2	Privacy By Design (PBD)	Oversight	Does the privacy policy contain self-reporting measures in case the firm experiences a privacy violation to, for example, a privacy seal organization or third party consultant?
S-1	Sharing (S)	privacy practices	Are affiliates or subsidiaries bound to this privacy policy, confidentiality agreements, or have contractual obligations outlining how the shared data will be used or secured?
S-1	Security (S)	Accuracy	Is there a term in the Privacy Policy or Terms of Use guaranteeing data accuracy?
S-2	Sharing (S)	privacy practices	Are contractors, service providers, or processors (for example, payment process companies) bound by either the same privacy policy, confidentiality agreements, or are under contractual obligations outlining how data will be used and secured?
S-2	Security (S)	Accuracy	Does the privacy policy specify any reasonable procedures the company may have in place to ensure data accuracy?
S-3	Sharing (S)	privacy practices	Are third parties bound by the same privacy policy?
S-3	Security (S)	Accuracy	Does the privacy policy note whether the firm reserves a right to disclose protected personally identifiable information to comply with law or prevent crime?
S-4	Sharing (S)	3rd Parties	Does the company perform due diligence to ensure the legitimacy of third parties that may have access to personally identifiable information?
S-4	Security (S)	Enforcement	Does the firm preserve the right to disclose protected identifiable information to protect its own rights?
S-5	Sharing (S)	3rd Parties	Does the company have a contract with third parties, other than processors or service providers, establishing how the shared personally identifiable data can be used?
S-5	Security (S)	Notice	Will users or consumers be given notice of any government requests for information about the user?

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

Question ID	Category	Subcategory	Question
S-6	Sharing (S)	3rd Parties	Does the privacy policy provide hyperlinks to the privacy policies of relevant third parties' PP's? For example, sometimes the privacy policy includes links to third party privacy policies when it states that any engagement with third parties will be governed by third party privacy policies
S-6	Security (S)	Notice	Does the privacy policy state whether the user will be notified in case there is a data breach?
S-7	Sharing (S)	Consent	What is consent mechanism for sharing or selling personally identifiable or sensitive information with entities that are not service providers? Please do not consider service providers whose function is to effect, administer, or enforce a transaction, send future correspondence to user, or perform research, internal database compilation, or servicing the website?
S-7	Security (S)	Privacy by Design	Does the privacy policy describe any substantive privacy and security protections incorporated into firm's managerial or structural procedures, such as limiting the number of employees who have access to personally identifiable data, allowing personally identifiable data access only for job-related functions, assigning employees to oversee privacy issues, employing Chief Privacy Officer, or requiring periodic audits?
S-8	Security (S)	Privacy by Design	Does the privacy policy identify which means of technological security it employs, such as encryption?
UC-1	User Control (UC)	Accuracy	Does the privacy policy allow the user or consumer to request that incorrect data be either rectified, updated, or erased?
UC-2	User Control (UC)	Accuracy	Does the privacy policy allow users or consumers to adjust their privacy settings? Note that directing the user to control cookies through settings in the browser doesn't count. The answer to this question may also be found by exploring the privacy settings of the service.
UC-3	User Control (UC)	Accuracy	Does the privacy policy allow users or consumers to access and correct or update any personally identifiable information collected by the company?
UC-4	User Control (UC)	User Control	Does the privacy policy allow the user to request that personally identifiable information be deleted or anonymized?

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

Question ID	Category	Subcategory	Question
UC-5	User Control (UC)	Ownership	Do the Terms of Use include a term explaining the ownership of the user's or consumer's personally identifiable information?
UC-6	User Control (UC)	Termination	Does the privacy policy state what happens to the data or personally identifiable information the company collects if the firm ceases to exist or is acquired?
UC-7	User Control (UC)	Termination	If the company is sold or goes bankrupt, is the user or consumer given choice as to what happens to their data or personally identifiable information?
UC-8	User Control (UC)	Termination	Does the privacy policy explain what happens to the personally identifiable information of a user who quits the service or closes the account?

The following performance measures reflect our proof-of-concept highlighter's ability to predict highlights.

<i>Question ID</i>	<i>Accuracy (sentences)</i>	<i>Recall (sentences)</i>	<i>f1 score (sentences)</i>	<i>Accuracy (paragraphs)</i>	<i>Recall (paragraphs)</i>	<i>f1 score (paragraphs)</i>
PP_in_TOU	0.80	0.78	0.45	0.86	0.85	0.47
v10.1_2020	0.76	0.76	0.44	0.82	0.78	0.46
v10.2_2020.1	0.84	0.81	0.46	0.87	0.88	0.48
v11.1_2020	0.80	0.63	0.45	0.86	0.68	0.47
v12_2020.1	0.74	0.76	0.44	0.76	0.77	0.45
v13_2020.1	0.73	0.72	0.43	0.80	0.75	0.46
v17_2020	0.78	0.81	0.45	0.84	0.83	0.47
v18_2020	0.73	0.74	0.43	0.77	0.76	0.46
v19_2020	0.81	0.78	0.45	0.84	0.76	0.46
v20_2020.1	0.77	0.77	0.44	0.85	0.75	0.47
v21_2020.1	0.80	0.78	0.45	0.88	0.87	0.48
v22_2020.1	0.80	0.72	0.45	0.85	0.77	0.47
v27_2020.1	0.88	0.80	0.47	0.93	0.83	0.49
v28_2020	0.85	0.84	0.46	0.91	0.85	0.49
v29_2020	0.84	0.78	0.46	0.86	0.79	0.47
v3_2020	0.67	0.62	0.40	0.73	0.65	0.43
v31_2020	0.75	0.71	0.43	0.81	0.76	0.46
v32_2020	0.87	0.81	0.48	0.92	0.88	0.51
v35_2020	0.91	0.81	0.48	0.96	0.87	0.52
v36_2020	0.89	0.77	0.47	0.95	0.89	0.50
v37.2_2020.1	0.81	0.78	0.45	0.83	0.83	0.47
v37_2020	0.81	0.80	0.45	0.86	0.83	0.47
v38_2020.1	0.78	0.75	0.45	0.80	0.81	0.45
v39_2020	0.79	0.78	0.45	0.83	0.86	0.47
v40_2020.1	0.83	0.76	0.46	0.91	0.89	0.49
v41_2020	0.88	0.84	0.47	0.92	0.73	0.48
v42_2020	0.78	0.68	0.44	0.84	0.74	0.46
v43_2020	0.86	0.81	0.47	0.88	0.87	0.49
v44_2020	0.86	0.81	0.47	0.90	0.85	0.49
v45_2020	0.88	0.80	0.47	0.90	0.80	0.48
v46_2020	0.84	0.82	0.46	0.89	0.79	0.48
v47_2020.1	0.86	0.86	0.47	0.91	0.85	0.49

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

<i>Question ID</i>	<i>Accuracy (sentences)</i>	<i>Recall (sentences)</i>	<i>f1 score (sentences)</i>	<i>Accuracy (paragraphs)</i>	<i>Recall (paragraphs)</i>	<i>f1 score (paragraphs)</i>
v48_2020	0.86	0.81	0.47	0.93	0.88	0.50
v49_2020	0.88	0.84	0.47	0.92	0.84	0.49
v51_2020	0.82	0.78	0.46	0.89	0.82	0.48
v52_2020	0.93	0.83	0.48	0.95	0.87	0.49
v53_2020	0.81	0.79	0.46	0.88	0.86	0.48
v54_2020	0.81	0.80	0.45	0.90	0.88	0.48
v55_2020	0.82	0.80	0.48	0.83	0.77	0.49
v56_2020	0.89	0.87	0.48	0.91	0.89	0.50
v57_2020	0.91	0.88	0.49	0.92	0.88	0.50
v58_2020	0.80	0.78	0.45	0.88	0.87	0.48
v59_2020	0.93	0.87	0.49	0.94	0.97	0.50
v60_2020	0.82	0.81	0.46	0.86	0.88	0.48
v61_2020	0.97	0.65	0.49	0.99	0.66	0.50
v62_2020.1	0.85	0.73	0.46	0.91	0.85	0.48
v63_2020	0.92	0.80	0.48	0.96	0.85	0.50
v64_2020	0.87	0.79	0.47	0.92	0.86	0.48
v71_2020.1	0.80	0.72	0.45	0.86	0.78	0.47
v72.1_2020.1	0.86	0.80	0.47	0.93	0.78	0.49
v72.2_2020.1	0.89	0.80	0.48	0.90	0.79	0.48
v72_2020.1	0.73	0.69	0.43	0.79	0.74	0.45
v73_2020.1	0.87	0.81	0.47	0.88	0.83	0.48
v74_2020.1	0.84	0.80	0.46	0.90	0.77	0.48
v75_2020.1	0.84	0.78	0.46	0.83	0.75	0.46
v76_2020.1	0.93	0.87	0.49	0.94	0.88	0.49
v77_2020	0.75	0.66	0.43	0.79	0.72	0.45
v78_2020.1	0.82	0.76	0.46	0.84	0.77	0.47
v80.2_2020.1	0.91	0.82	0.48	0.96	0.73	0.50
v80.3_2020.1	0.89	0.80	0.47	0.90	0.78	0.48
v80_2020.1	0.85	0.81	0.46	0.91	0.92	0.49
v81_2020	0.87	0.76	0.47	0.92	0.73	0.48
v82_2020	0.96	0.79	0.50	0.97	0.99	0.50
v83_2020	0.97	0.79	0.50	0.98	0.91	0.50

Appendix _: Automate Coding

A. Regression model using embeddings from generated excerpts

The following table shows how our tuned coder performs at predicting the J.D. student coder’s responses.

1. Coding Predictions

Question ID	Coders & AI match	Coders disagree; AI agrees with one coder	AI does not match any coder	AI predicts at least one coder	In the case of disagreement, AI is the odd one out
PP_in_TOU	71%	20%	9%	91%	16.67%
v10.1_2020	78%	13%	9%	91%	22.92%
v10.2_2020.1	57%	11%	31%	69%	23.64%
v11.1_2020	46%	23%	31%	69%	17.33%
v12_2020.1	55%	27%	17%	83%	19.77%
v13_2020.1	61%	19%	20%	80%	18.57%
v17_2020	66%	8%	27%	73%	27.69%
v18_2020	43%	28%	29%	71%	23.38%
v19_2020	75%	8%	18%	82%	27.42%
v20_2020.1	46%	17%	37%	63%	19.05%
v21_2020.1	58%	14%	29%	71%	20.97%
v22_2020.1	46%	15%	39%	61%	23.47%
v27_2020.1	73%	14%	13%	87%	16.22%
v28_2020	38%	16%	46%	54%	26.96%
v29_2020	75%	9%	16%	84%	24.07%
v31_2020	37%	23%	41%	59%	29.90%
v32_2020	91%	4%	5%	95%	26.92%
v35_2020	75%	8%	17%	83%	25.00%
v36_2020	82%	12%	6%	94%	20.69%
v37.2_2020.1	61%	11%	29%	71%	21.28%
v37_2020	69%	20%	11%	89%	17.86%
v38_2020.1	66%	10%	23%	77%	30.51%
v39_2020	48%	21%	31%	69%	20.65%
v3_2020	79%	13%	7%	93%	20.83%
v40_2020.1	63%	10%	26%	74%	22.58%
v41_2020	92%	6%	2%	98%	18.75%
v42_2020	68%	23%	9%	91%	18.64%
v43_2020	95%	2%	4%	96%	29.41%
v44_2020	84%	7%	9%	91%	25.00%
v45_2020	70%	18%	12%	88%	28.79%
v46_2020	77%	8%	15%	85%	30.77%
v47_2020.1	54%	11%	34%	66%	28.38%
v48_2020	64%	20%	16%	84%	25.00%
v49_2020	67%	21%	12%	88%	13.33%
v51_2020	46%	16%	38%	63%	27.91%
v52_2020	72%	17%	11%	89%	22.58%
v53_2020	68%	21%	11%	89%	21.88%
v54_2020	81%	5%	15%	85%	30.36%
v55_2020	37%	28%	35%	65%	31.72%
v56_2020	75%	14%	11%	89%	24.56%
v57_2020	84%	5%	10%	90%	27.27%
v58_2020	65%	6%	29%	71%	29.49%

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

<i>Question ID</i>	Coders & AI match	Coders disagree; AI agrees with one coder	AI does not match any coder	AI predicts at least one coder	In the case of disagreement, AI is the odd one out
v59_2020	82%	3%	15%	85%	31.58%
v60_2020	69%	19%	12%	88%	21.54%
v61_2020	96%	3%	1%	99%	14.29%
v62_2020.1	85%	11%	4%	96%	14.29%
v63_2020	76%	13%	12%	88%	26.53%
v64_2020	81%	9%	10%	90%	30.00%
v71_2020.1	78%	10%	12%	88%	26.42%
v72.1_2020.1	65%	9%	25%	75%	27.59%
v72.2_2020.1	57%	13%	30%	70%	25.71%
v72_2020.1	21%	5%	75%	25%	27.06%

2. *Confidence predictions*

APPENDIX TABLE 1: Labelling Scheme

APPENDIX _: RATIONALE FOR BUILDING A CUSTOM CODING TOOL

Our decision to build a custom coding tool was motivated by several rationales.

First, an online coding tool improved data consistency. Automated systems work best with unambiguously formatted, machine-readable data. Even seemingly inconsequential variances, such as inconsistent punctuation in variable names, can create the need for a great deal of time-consuming “data cleaning” to prepare a dataset for use in training machine learning algorithms or conducting other computer-assisted processing. The risk of inconsistent coding approaches was magnified by the pandemic and academic calendar; our research assistants were scattered across time zones, sometimes had inconsistent internet connectivity, and worked asynchronously over a period of years. Adding additional technical training and feedback on formatting and managing the maintenance of shared documents would complicate the already substantial training involved in teaching RAs to code documents. Online tools provide an easy solution to this problem.

Second, custom coding software allowed us to collect more information. For example, for each question and document collection, we recorded timing information, self-reported confidence, had RAs highlight the text relevant to the question, and allowed RAs to add comments and notes.

Third, the extra data and ability to adjust the coding method programmatically allowed us to iteratively improve our methodological design. For example, we used the first three months of the project to iteratively improve our coding questions. When we adjusted or added questions, DocumentCoder automatically removed outdated answers from completed codings and added new or changed questions to RAs’ task lists. When we removed questions, DocumentCoder stopped showing those questions. As discussed above, the ability to review and discuss disagreements allowed us to avoid creating ambiguity in our classification, improving our ability to find ambiguity in the contract.

Finally, DocumentCoder is freely available to researchers. We hope it will provide three benefits. First, a shared coding tool allows overlapping research efforts to more easily pool or compare their data. For example, other projects studying privacy policies could use our tool to extend our dataset and uncover new comparative insights without re-coding any English-language policies from 2020. Second, a shared coding tool can help empirical legal scholars to leverage innovation in computer science and machine learning without needing to augment their legal and statistical expertise with up-to-date programming skills and AI knowledge. Automated analysis of legal documents requires disparate skills; knowledge of a specific, relevant area of law; expertise in data science and statistics, and the ability to write sophisticated computer programs. Programs written to analyze one dataset prepared using DocumentCoder will work on any dataset prepared using DocumentCoder, allowing legal experts to reproduce analytical approaches without learning to program or finding a software engineer collaborator (programs that use DocumentCoder’s plugin API can be installed like an app). Third, a shared coding tool can help foster interdisciplinary coordination. If an AI researcher writes an algorithm that uses our dataset to train an AI to read privacy policies, that algorithm should also work for other collections of documents—like caselaw or articles of incorporation—coded using the tool. This can have a positive feedback effect: making AI results more broadly applicable to legal scholars increases the potential impact of that research, and using a format that is already compatible with existing

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

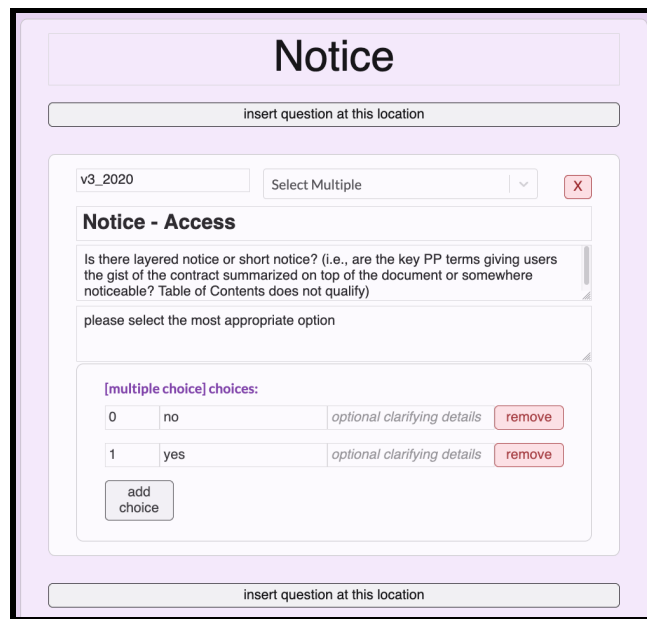
ML tools unlocks immediate research benefits for empirical legal researchers coding new document troves.

B. Core Features and Functional Walkthrough of the DocumentCoder

We designed DocumentCoder to capture the legal interpretations of documents in a machine-readable format that could train machine-learning algorithms. We included support for labeled phrases, designed mechanisms to detect proxies of potential ambiguity, and recorded holistic codings to support future algorithmic work—a bet that is paying off following the advent of “transformer” AIs that can read thousands of words in a single pass.

Document and Metadata Upload. Documents can be uploaded to DocumentCoder manually or through an API. We set up the API to support our own web scraper and to connect to databases like the Princeton-Leuven Corpus. DocumentCoder can also parse most PDFs and scrape most URLs. Uploaded documents are labeled with a source (e.g., “Google’s privacy policy, terms of service, and other documents incorporated by reference”) and a snapshot (e.g., “scraped from Google’s U.S. webpage on June 7, 2020”). A source might contain multiple snapshots—the same policy captured at different times or from different contexts, for example. During the coding process, snapshots are presented to coders holistically. That is, a coder can be presented with multiple interrelated documents at once.

Question Creation and Management. DocumentCoder also contains a tool for managing the questions presented to coders. Questions are multiple choice, and can be set to allow one or many responses. Questions can be clustered into groups, which allows the tool to track timing metrics across related questions.



The screenshot displays a web interface for creating a question. At the top, there is a header labeled "Notice". Below it, a button says "insert question at this location". The main content area shows a question being created for a document labeled "v3_2020". The question title is "Notice - Access" and the text is "Is there layered notice or short notice? (i.e., are the key PP terms giving users the gist of the contract summarized on top of the document or somewhere noticeable? Table of Contents does not qualify)". Below the question text, there is a prompt "please select the most appropriate option". Underneath, there is a section for "[multiple choice] choices:" with two options: "0 no" and "1 yes". Each option has a text input field for "optional clarifying details" and a "remove" button. At the bottom of the choices section, there is an "add choice" button. The interface also features a second "insert question at this location" button at the very bottom.

Coding Process. Coders are assigned groups of related documents and a list of questions, which are displayed in a split-screen interface. The right pane displays the document, the left pane the list of questions. Coders choose a question, highlight

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

the pertinent text within the document, answer the question, indicate their confidence level, and optionally add comments. DocumentCoder records this data, along with the time taken to answer each question.

If multiple coders code the same set of documents, DocumentCoder collates those responses. If questions change, DocumentCoder preserves the responses to unchanged questions while marking the altered question as incomplete.

The screenshot displays the DocumentCoder interface. On the left, a document titled "DOCUMENT A - Amazon.com Help: Amazon.com Privacy Notice" is open. The document text includes instructions for configuring settings and providing information. On the right, a "Notice" pane shows a question: "Does the company explicitly state they use tracking elements other than cookies? (e.g. 'local storage cookies', 'browser fingerprints')?". The question is marked with a confidence level of "very high" and a count of "1" sentence. Below the question, there are radio buttons for "no", "yes", "does not disclose", and "other (enter here)". The "yes" option is selected. A confidence scale is shown with buttons for "very low", "low", "medium", "high", and "very high". At the bottom of the interface, there are navigation buttons: "A B C D E F", "scroll to: Current Question", "Next Unanswered", and "Save and return home".

Review Process. Reviewers use a similar split-screen interface as coders, but are shown the coders response in the the questions pane. The interface flags locations where coders disagree on answers or highlighting, report low confidence, or left comments. Reviewers have the option to answer questions, provide comments, and indicate confidence level in the same manner as coders. When collating answers, DocumentCoder treats reviewer responses as authoritative when provided.

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

DOCUMENT A - Amazon.com Help: Amazon.com Privacy Notice top

history to your account, you may do so by logging out of your account [here](#)[7] and blocking cookies on your browser.

- * You can manage the recommendations you receive in our store [here](#)[11], remove recommendations you don't want to see [here](#)[12] by selecting View All and Manage then selecting the Remove Items toggle that appears at the top of the page, and edit your browsing history [here](#)[13].
- * You will also be able to opt out of certain other types of data usage by updating your settings on the applicable Amazon website (e.g., in "Manage Your Content and Devices"), device, or application. For more information click [here](#)[14]. Most non-Amazon devices also provide users with the ability to change device permissions (e.g., disable/access location services, contacts). For most devices, these controls are located in the device's settings menu. If you have questions about how to change your device permissions on devices manufactured by third parties, we recommend you contact your mobile service carrier or your device manufacturer.
- * If you are a seller, you can add or update certain information in [Seller Central](#)[15], update your account information by accessing your [Seller Account Information](#)[16], and adjust your email or other communications you receive from us by updating your [Notification Preferences](#)[17].
- * If you are an author, you can add or update the information you have provided in the [Author Portal](#)[18] and [Author Central](#)[19] by accessing your accounts in the Author Portal and Author Central, respectively.

27 In addition, to the extent required by applicable law, you may have the right to request access to or delete your personal information. If you wish to do any of these things, please contact [Customer Service](#)[9]. Depending on your data choices, certain services may be limited or unavailable.

28 **[\$1. 9] Are Children Allowed to Use Amazon Services?**

29 Amazon does not sell products for purchase by children. We sell children's products for purchase by adults. If you are under 18, you may use Amazon Services only with the involvement of a parent or guardian. We do not knowingly collect personal information from children under the age of 13 without the consent of the child's parent or guardian. For more information, please see our [Children's Privacy Disclosure](#)[20].

30 **[\$1. 10] EU-US and Swiss-US Privacy Shield**

Amazon.com Inc. participates in the EU-US and Swiss-US Privacy Shield

A B C D E F

CCPA

10. CCPA - Opt Out (v80_2020.1)

answers do not match	sentences do not match
cv729@nyu.edu, kat9234@nyu.edu, gfc9001@nyu.edu, florenca.m.wurgler@gmail.com	

Firm offers the right of opt-out of selling personal information to third parties with a visible, direct link to "Do Not Sell My Personal Information."

please select the most appropriate option

cv729@nyu.edu	2	0 sentences flagged	confidence: 4
kat9234@nyu.edu	2	1 sentences flagged	confidence: 4
gfc9001@nyu.edu	0	4 sentences flagged	confidence: 4
<ul style="list-style-type: none"> ¶ A.21(10) ¶ A.26(13) ¶ B.12(1) ¶ F.7(2) 			

Reviewer response

florenca.m.wurgler@gmail.com 2 1 sentences flagged

confidence: unspecified

¶ A.14(1)

Sentences highlighted by everyone:

no mention

yes

mentions the right but explains why it's not available/not applicable (e.g., firm does not sell information to third parties)

N/A

other (enter here)

confidence:

very low

low

medium

high

very high

additional comments

Progress:

scroll to: Current Question Next Unresolved Save and return home

Project Tracking. To help keep track of progress across our disparate community of RAs, we build several project tracking features into DocumentCoder. A project dashboard shows the number of questions left to answer for each coder on each document. It also tracks which documents are fully coded but pending review. These dashboards are highly reconfigurable through an API, a feature we added in order to track progress on new or changed questions separately from our unchanged labels.

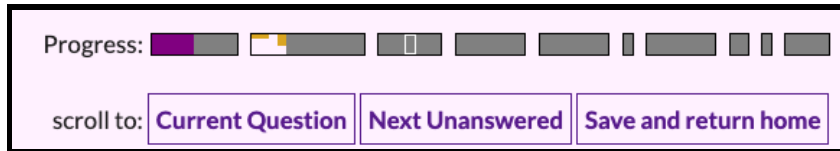
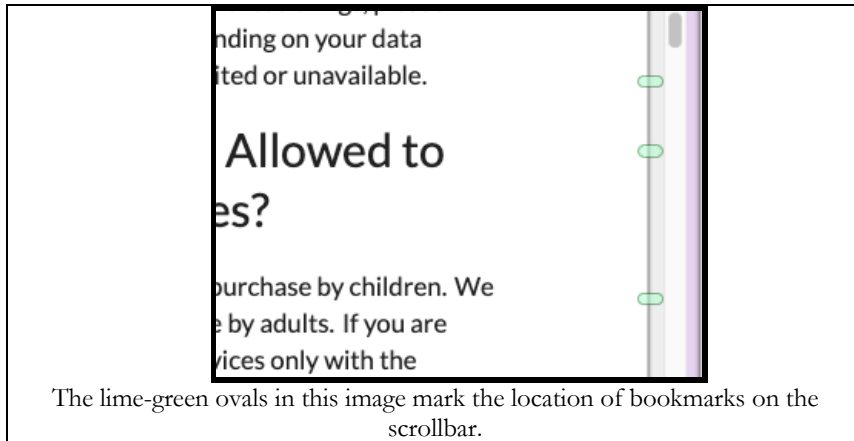
Policy	Docs Loaded	Coders Done	Fully Reviewed	links
aboutus.disaboom.com	done	✓	🔄	Code Review
webmediabrands.com	done	✓	🔄	Code Review
dictionary.com	done	✓	🔄	Code Review
amazon.co.uk	done	✓	🔄	Code Review
sediabio.com	done	✓	🔄	Code Review

API and Data Export. We added APIs and data export features to compute summary statistics, visualization data, and train a proof-of-concept machine learning model, discussed further below.

C. *Development Approach*

We adopted an agile approach to tool development, starting with a minimally viable product and iterating continuously until we had a fairly stable and usable tool. After building the initial version of DocumentCoder in just under two weeks, we immediately began coding documents. Over the first two months of the project, we updated the tool and our questions based on insights gleaned from weekly, hour-long review meetings with each coder. We adjusted the tool with the goal of allowing coders to spend as much time as possible reading, and as little time as possible wrestling with the software.

We found that student coders quickly become “power users,” preferring speed and information density over discoverability. In response to coder feedback, we added several quality-of-life improvements to the coding screen. The first things we added were multiple keyboard shortcuts for every action. We then added a plethora of navigation tools, including question status bar, a “scroll to next unanswered question” button, a way for coders to “bookmark” and fast scroll to paragraphs, grouping questions into categories and allowing bookmarks to only appear for specific questions or categories. We also added features to highlight changed questions, and tools for coders to view pending assignments and browse to their assigned snapshots.



We also addressed several unanticipated error modes as they came up. Our main source of trouble came from the many ways connectivity problems interact with online tools. We added support for offline work, time zone issues (for a coder who worked in offline mode during flights), and we also built mechanisms for saving and restoring local copies of codings. We addressed conflicts arising from coders opening the same snapshot in multiple tabs or multiple computers.

The first version of DocumentCoder supported the more granular labeling approach used in the OPP-115 dataset, complete with a reproduction of their labeling tool. Running the OPP labels through our review process consumed the lion's share of our review time over the first few months, and we found that most disagreements revolved around ambiguity in the text of the questions, rather than ambiguity in the text of the contract. Because DocumentCoder shows the entire collection of documents at once instead of showing about a paragraph's worth of text at a time, which created confusion and frustration for coders. We chose to cut the questions and remove our reproduction of the OPP coding tool during the initial iterative stage of the project.

Bibliography

Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., & Mayer, J. (2021, April). Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021* (pp. 2165-2176).

Blair-Stanek, Andrew, Nils Holzenberger, and Benjamin Van Durme. "Can GPT-3 perform statutory reasoning?" In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pp. 22-31. 2023.

Blair-Stanek, Andrew, Anne-Marie Carstens, Daniel S. Goldberg, Mark Graber, David C. Gray, and Maxwell L. Stearns. "GPT-4's Law School Grades: Con Law C, Crim C-, Law & Econ C, Partnership Tax B, Property B-, Tax B." *Crim C-, Law & Econ C, Partnership Tax B, Property B-, Tax B* (May 9, 2023) (2023).

Choi, Jonathan H., Kristin E. Hickman, Amy B. Monahan, and Daniel Schwarcz. "ChatGPT goes to law school." *J. Legal Educ.* 71 (2021): 387.

Dahl, Matthew, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. "Large legal fictions: Profiling legal hallucinations in large language models." *arXiv preprint arXiv:2401.01301* (2024).

Jens Frankenreiter and Julian Nyarko. *Natural language processing in legal tech. Legal Tech and the Future of Civil Justice* (David Engstrom ed.), 2022.

Guha, Neel, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters et al. "Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models." *Advances in Neural Information Processing Systems* 36 (2024).

Harkous, H., Fawaz, K., Lebre, R., Schaub, F., Shin, K. G., & Aberer, K. (2018). *Polisis: Automated analysis and presentation of privacy policies using deep learning*. In *27th USENIX Security Symposium (USENIX Security 18)* (pp. 531-548).

Linden, T., Khandelwal, R., Harkous, H., & Fawaz, K. (2018). The privacy policy landscape after the GDPR. *arXiv preprint arXiv:1809.08396*; Fawaz, K., Linden, T., & Harkous, H. (2019, January). The applications of machine learning in privacy notice and choice. In *2019 11th International Conference on Communication Systems & Networks (COMSNETS)* (pp. 118-124). IEEE

Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H.W., Sartor, G. and Torroni, P., 2019. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27, pp.117-139.

Mousavi Nejad, N., Jabat, P., Nedelchev, R., Scerri, S. and Graux, D., 2020. Establishing a strong baseline for privacy policy classification. In *ICT Systems Security and Privacy Protection: 35th IFIP TC 11 International Conference, SEC 2020, Maribor, Slovenia, September 21–23, 2020, Proceedings 35* (pp. 370-383). Springer International Publishing.

Nejad, N. M., Graux, D., & Collarana, D. (2019, June). *Towards Measuring Risk Factors in Privacy Policies*. In *ALAS@ ICAIL* (pp. 18-20); Mousavi Nejad, N., Jabat, P., Nedelchev, R., Scerri, S., & Graux, D. (2020). Establishing a strong baseline for privacy policy classification. In *ICT Systems Security and Privacy Protection: 35th IFIP TC 11 International Conference, SEC 2020, Maribor, Slovenia, September 21–23, 2020, Proceedings 35* (pp. 370-383). Springer International Publishing

Okoyomon, E., Samarin, N., Wijesekera, P., Elazari Bar On, A., Vallina-Rodriguez, N., Reyes, I., ... & Egelman, S. (2019, May). *On the ridiculousness of notice and consent: Contradictions in app privacy policies*. In Workshop on Technology and Consumer Protection (ConPro 2019), in conjunction with the 39th IEEE Symposium on Security and Privacy.

Thalke, Rosamond, Edward H. Stiglitz, David Mimno, and Matthew Wilkens. "Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement." *arXiv preprint arXiv:2310.18440* (2023).

Zaeem, R. N., & Barber, K. S. (2021). Comparing Privacy Policies of Government Agencies and Companies: A Study using Machine-learning-based Privacy Policy Analysis Tools. In *ICAART (2)* (pp. 29-40), and sources cited therein.

CAN LLMs READ PRIVACY POLICIES AS WELL AS LAWYERS?

Zheng, L., Guha, N., Anderson, B.R., Henderson, P. and Ho, D.E., 2021, June. 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 (pp. 159-168).