Building a Long Text Privacy Policy Corpus with Multi-Class Labels

Florencia Marotta-Wurgler[*]
NYU School of Law
wurglerf@exchange.law.nyu.edu

David Stein[*]
Northeastern School of Law
Khoury College of Computer Sciences
da.stein@northeastern.edu

Abstract

This work introduces a new hand-coded dataset for the interpretation of privacy policies. The dataset captures the contents of 162 privacy policies, including documents they incorporate by reference, on 64 dimensions that map onto commonly found terms and applicable legal rules. The coding approach is designed to capture complexities inherent to the task of legal interpretation that are not present in current privacy policy datasets. These include addressing textual ambiguity, indeterminate meaning, interdependent clauses, contractual silence, and the effect of legal defaults.

1 Introduction

Privacy policies are important legal documents; they govern how firms collect, use, share, and secure personal information. They have become a prime target for automated interpretation. Despite their importance, they are rarely read. Privacy policies are long, complex, and require legal expertise to understand. At the same time, many privacy policies are publicly available and map to a consistent set of well-defined legal questions. This presents an opportunity to represent policy content in a reasonably consistent and somewhat objective manner against which automated interpreters can be tuned and measured.

Recent advancements in machine learning, especially the introduction of the large language model (LLM), increased interest in automating long-text interpretation. Practical legal use of NLP has expanded from term- and clause-level classification to nuanced interpretation of long bodies of legal text. Legal interpretation presents an especially challenging interpretative task. Privacy policies are a stereotypical legal document, as they are drafted by experts and include domain-specific vocabulary and interpretation (Zheng et al., 2021; Mellinkoff, 2004; Mertz, 2007); they can contain inconsistencies and be susceptible to multiple valid interpretations (Reidenberg et al., 2016); and they often contain interdependent clauses—sometime spread across multiple documents—whose meaning is best understood when read wholesale. Like other legal texts, privacy policies must be interpreted in the context of applicable legal rules, which can define terms and provide guidance on issues not explicitly addressed in the text.

Current privacy policy datasets either offer high-granularity labels for short samples of policy text, or low-granularity classification of longer text. These approaches may not capture many domain-specific aspects of legal interpretation that are relevant to the expanding range of automated legal tasks. For example, neither approach accounts for how documents that are "incorporated by reference" may affect the way a policy restricts (or doesn't restrict) the ways in which a company can use user data. As legal interpretation increasingly becomes the target of automation, new datasets are needed. This paper aims to help address that need.

We provide a legal dataset of labeled online privacy policies coded by legally-trained experts. It contains 162 privacy policies along with the documents they incorporate by reference, including Terms of Use, Cookie Policies, California Consumers Privacy Act (CCPA) disclosures, and documents pertaining to compliance with the European Union's General Data Protection Regulation (GDPR). Our coding accounts for the ways in which applicable legal rules and referenced documents can affect the meaning of terms. It also tracks relevant legal terms and reflects aspects of the legal interpretation task, including accounting for ambiguity and reasonable disagreements, and interpreting silence.

[*] Equal contribution

2 Related Work

Prior work building datasets for privacy policies mostly focuses on expert annotation or classification of short text. (Lippi et al., 2019; Bui et al., 2021; Ahmad et al., 2021). Some investigation has also looked into crowd-sourcing annotation (Wilson et al., 2018). One privacy-policy-adjacent dataset involving classification of longer legal text labels the content of cookie banner disclosures with the stated purposes for data collection (Santos et al., 2021). In addition to annotated datasets, there are large-scale compilations of privacy policies scraped from the Internet and Internet Archive(Amos et al., 2021; Srinath et al., 2021).

Perhaps the most widely-used privacy policy dataset is the OPP-115 dataeset introduced by Wilson et. al. in 2016. OPP contains 115 privacy policies that were annotated paragraph-by-paragraph to identify phrases related to 36 data practices grouped into 10 categories. The OPP dataset was used to train prominent tools used to pick out specific clauses from privacy policies (Harkous et al., 2018; Mousavi Nejad et al., 2020). It has also been used to generate related datasets, either by transforming its annotations for use in a new task like question-answering or GDPR compliance (Poplavska et al., 2020; Ahmad et al., 2020), or as an input into composite legal-task benchmarks like LEGALBENCH and PRIVACYGLUE (Guha et al., 2023; Chalkidis et al., 2022). The OPP taxonomy scheme has also been used to organize other privacy-related datasets (Ravichander et al., 2019). Another notable privacy policy dataset—the *unfair-TOS* dataset introduced by Lippi et. al.—annotates "potentially unfair" clauses in privacy practices and is also incorporated into some composite benchmarks, including the privacy-policy-specific PRIVACYGLUE benchmark (Shankar et al., 2023).

Benchmarking legal AIs goes beyond the traditional metrics-and-datasets approach. Alternative evaluation approaches include having NLP systems take the bar exam (Bommarito II and Katz, 2022; Katz et al., 2024) (though some have questioned the efficacy of that evaluation approach (Martínez, 2024)), grading LLM-generated law school exam answers (Choi et al., 2021), and measuring how law student performance is affected by LLM use (Choi and Schwarcz, 2023).

3 Dataset Preparation

3.1 Document Selection

Privacy policies are the legal documents that govern the relationship between data collectors (usually firms) and data subjects (usually consumers) regarding the collection, use, sharing, and security of their personally identifiable information. They create rights and obligations and comply with legal requirements, such as the Children Online Privacy Protection Act (COPPA). Most privacy policies are posted online on companies' websites. They "incorporate by reference" other documents with additional terms that also govern the user/firm relationship, such as Terms of Use, Cookie Policies, CCPA, GDPR disclosures. For example, all terms related to tracking may only appear in a linked Cookie Policy document.

Taken together, this constellation of documents comprises the scope of terms on personal information privacy. Our approach focuses on collecting the set of all interrelated documents. For each website in our sample, we collect the privacy policy, Terms of Use (if available), documents incorporated by reference (either directly or indirectly), and documents referred to in connection to an “I agree” button in an account-creation flow (if applicable).

Privacy policies vary significantly between services. Prior research reported statistically significant differences in the content across markets and website popularity (Marotta-Wurgler, 2016b). Our dataset includes samples distributed roughly logarithmically by traffic rank and includes examples from each tier-1 industry category in the IAB Content Taxonomy 2.0 (IAB Tech Lab, 2024). We compiled our sample by selecting from websites bucketed by order-of-magnitude rank. If a bucket contained websites from prior privacy policy datasets (Marotta-Wurgler, 2016b; Wilson et al., 2016), we selected randomly from those previously encoded websites. We selected randomly from the relevant rank bucket in the few cases where no previously encoded side was available. Our initial selection process contained representatives from all industry categories; we did not re-sample to adjust coverage. Figure 1 shows the distribution of policies collected.

3.2 Coding Process

Each policy was coded independently by two law students who had completed relevant coursework

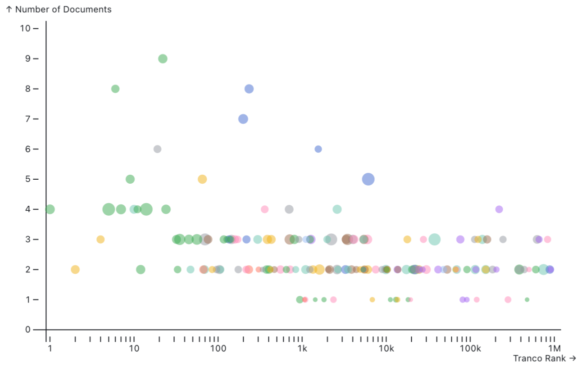


Figure 1: The documents in the dataset. The X axis shows the website's Tranco traffic rank. The Y axis reflects the number of documents collected for that site. Color shows the website's industry category. Radius corresponds to total word count. An 18-document outlier is omitted from this chart.

on contracts and received training regarding legal interpretation of privacy policies. We presented each coder with all relevant legal documents corresponding to a firms' website and multiple-choice questions about those documents. For each question, coders were asked to highlight any and all text (including text in documents incorporated by reference) they found relevant to answering the question. They answered each question by selecting from a set of choices describing possible policy content, including, when applicable, the possibility that the policy was silent on a question. They also recorded their confidence in their answer on a Likert scale.

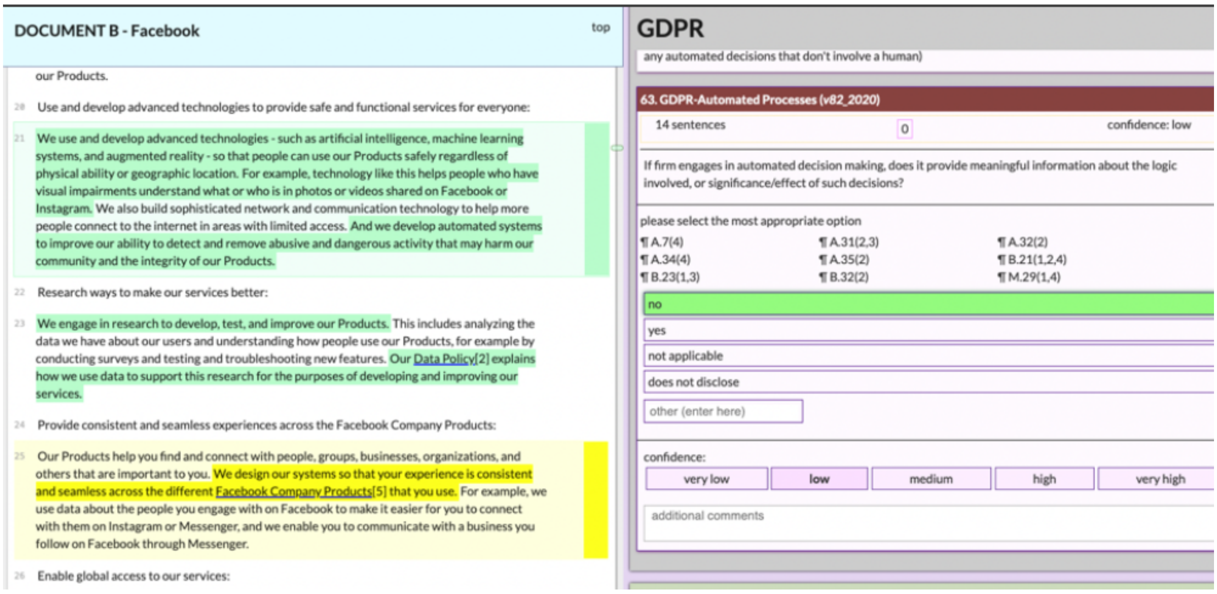
We reviewed the entries of each coder, resolving disagreements and coding errors where appropriate, as described in detail in the next subsection. For every question where we resolved disagreements or corrected errors, we also highlighted text we found relevant, recorded our answer, and rated our confidence. Our responses are included in the complete dataset. They are not used in this paper's summary statistics or benchmarks.

Coding, review, and project management were all performed using a suite of custom web tools we developed, as shown in figure 2. We provide a hosted version of the tools on our website. The source code is available on our GitHub repository and a hosted version of the tool is available at documentcoding.com.

3.3 Coding Schema

We generated our coding schema following the procedure developed by Marotta-Wurgler (Marotta-Wurgler, 2016a). The approach has been used in legal empirical scholarship to make quantitative comparisons of privacy policy content and compliance between industries and over time (Marotta-Wurgler, 2016b,c; Davis and Marotta-Wurgler, 2019). The schema represents a policy content as a set of labels derived from significant and influential privacy guidelines and applicable rules that have shaped the content and structure of privacy policies. These are: the 1973 HEW Fair Information Practice Principles, the 2012 Federal Trade Commission's Information Privacy Guidelines, the 2012 White House Privacy Bill of Rights, the GDPR of 2018, and the CCPA of 2020, and contract law—the background rules courts have employed to enforce privacy policy. The resulting coding provides a granular representation of privacy policy content mentioned in relevant guidelines and laws. For example, there are three labels that encode the rights users and firms have with respect to changes to the policy (can the firm make changes, does a user have to assent to that change before it takes effect, and are changes retroactive). Another label marks whether the set of documents includes a class action waiver. The goal of a granular approach was to minimize ambiguity in representation of policy content and enhance consistency among coders. We translate these variables into 64 multi-choice questions, which we group into 11 categories:

1. *CCPA* (10 labels): Tracks requirements unique to the California Consumer Protection Act, such as whether the subject can request that their personal information not be sold.
2. *GDPR* (5 labels): Tracks requirements unique to the GDPR, such as whether the entity has designated a Data Privacy Officer.
3. *Data Practices (DP)* (1 label): whether the firm has procedures to safely dispose of personal information.
4. *Enforcement (E)* (8 labels): Tracks mechanisms of legal reddress.
5. *Notice (N)* (13 labels): Tracks notices pertaining to data collection and mandatory disclosures with state privacy laws.
6. *Contract (K)* (1 label): Tracks whether policy incorporates terms by reference.
7. *Privacy by Design (PBD)* (2 labels): Tracks general data practices in management and design.
8. *Security (SE)* (8 labels): Tracks information security practices.



DOCUMENT B - Facebook

our Products.

20 Use and develop advanced technologies to provide safe and functional services for everyone:

21 We use and develop advanced technologies - such as artificial intelligence, machine learning systems, and augmented reality - so that people can use our Products safely regardless of physical ability or geographic location. For example, technology like this helps people who have visual impairments understand what or who is in photos or videos shared on Facebook or Instagram. We also build sophisticated network and communication technology to help more people connect to the internet in areas with limited access. And we develop automated systems to improve our ability to detect and remove abusive and dangerous activity that may harm our community and the integrity of our Products.

22 Research ways to make our services better:

23 We engage in research to develop, test, and improve our Products. This includes analyzing the data we have about our users and understanding how people use our Products, for example by conducting surveys and testing and troubleshooting new features. Our Data Policy[2] explains how we use data to support this research for the purposes of developing and improving our services.

24 Provide consistent and seamless experiences across the Facebook Company Products:

25 Our Products help you find and connect with people, groups, businesses, organizations, and others that are important to you. We design our systems so that your experience is consistent and seamless across the different Facebook Company Products[5] that you use. For example, we use data about the people you engage with on Facebook to make it easier for you to connect with them on Instagram or Messenger, and we enable you to communicate with a business you follow on Facebook through Messenger.

26 Enable global access to our services:

GDPR

any automated decisions that don't involve a human)

63. GDPR-Automated Processes (v82_2020)

14 sentences 0 confidence: low

If firm engages in automated decision making, does it provide meaningful information about the logic involved, or significance/effect of such decisions?

please select the most appropriate option

¶ A.7(4) ¶ A.31(2,3) ¶ A.32(2)
¶ A.34(4) ¶ A.35(2) ¶ B.21(1,2,4)
¶ B.23(1,3) ¶ B.32(2) ¶ M.29(1,4)

no

yes

not applicable

does not disclose

other (enter here)

confidence:

additional comments

Figure 2: The highlight tool

9. *Sharing (SH)* (7 labels): Tracks sharing with third and other parties.
10. *User Control (UC)* (8 labels): Tracks user rights regarding personal information access and control.

For each question we drafted a set of answers designed to minimize ambiguity while providing granular representation of policy content. These include choice sets that are binary (*"Do the Terms of Use or Terms of Service incorporate the Privacy Policy by reference?" [yes/no]*), single-class (*"Does the Privacy Policy offer data requests by consumers explicitly free of charge?" [Not Applicable/Yes/No]*), and multi-class (*"Does the privacy policy provide means by which a user can contact the company with any privacy concerns or complaints? [select all that apply...]"*). Response options also include and distinguish between policies being silent regarding a term or the particular term being not applicable (e.g., a policy that states that no personal information is collected does not need to provide information about how such information is stored).

In contrast to other privacy policy datasets, which make efforts to maximize inter-coder agreement and often discard points of disagreement, we preserve disagreement and low-confidence coding. Because ambiguity is feature of many legal texts, the ground truth is effectively probabilistic, meaning disagreement and low confidence are expected features of a classification task that capture a particular nuance of the task of legal interpretation. To ensure that disagreements and reported low confidence correspond to ambiguity in the policy rather than unclear coding instructions, we engaged in an iterative process to reduce exogenous sources of ambiguity from our coding.

Once a week during the 10-week iterative revision period, we met with coders and discussed each of their coding choices. The recorded highlights for each question helped coders recall and explain their decision-making. We qualitatively assessed the source of each instance of inter-coder disagreement or or low reported confidence, choosing between five possible causes:

1. *Questions*: coders interpret the question in conflicting ways due to poor or confusing wording
2. *Choices*: the answer choice sets did not fully map onto the text and law
3. *Defaults*: the answer choice set does not properly account for the existence of default rules that alter the meaning of contractual silence
4. *Coder Error*: a coder made a mistake.
5. *Policy Text*: the text of the policy is ambiguous or susceptible to reasonable disagreements in interpretation

We addressed disagreements or confusion resulting from Categories 1, 2, and 3 by either adding clarifying details to the language of questions and answer choice sets, or adjusting the set of choices to better reflect the range of observed practices.

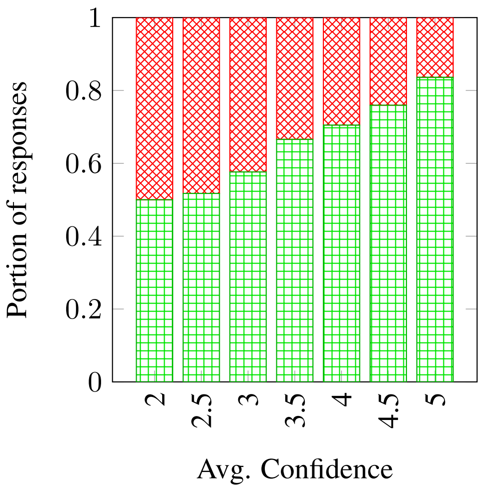


Figure 3: Agreement rate among coders, by average self-reported confidence on a Likert scale.

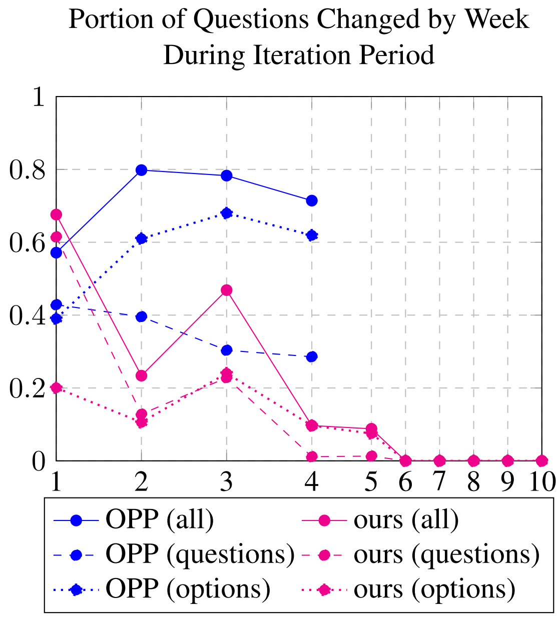


Figure 4: Portion of questions changed during iterative refinement, by coding scheme. OPP was removed from the scheme after week 5.

When we detected coder error we corrected it and send updated training guidance to other coders. We made no changes following disagreements and low confidence caused by unclear policy text. After any change to a question we removed any coding recorded using an outdated version of that question from the dataset. Coders relabeled those questions at the end of the iterative adjustment period.

Not every policy implicates the same questions or choices; some issues arose later in our revision window. After five rounds of iteration and revision, we stopped observing instances of the first three categories. The final set of coding instructions, including the history of changes corresponding to each variable, is included as an online appendix.

As an initial sanity check on our data, we compare confidence and inter-coder agreement rates, as shown in figure 3. We observe that inter-coder agreement has a roughly linear relationship with self-reported confidence. This suggests that both disagreement and confidence correspond with situations where coders see multiple potentially appropriate answers. Our efforts to remove other sources of confusion and our use of expert coders mean that this ambiguity should largely come from the text of the policy.

3.4 Insights from Iterative Schema Refinement

To test whether our coding scheme and iterative refinements actually reduce measurement errors by reducing exogenous sources of ambiguity, we included the OPP annotation scheme for the first five weeks of our iterative process. OPP is a natural baseline to measure against: it is cited as the current "gold standard" privacy policy annotation scheme (Mousavi Nejad et al., 2020), and has been used to train numerous automated privacy policy interpreters and included in the LegalBench and PrivacyGLUE composite legal reasoning benchmarks. While the OPP taxonomy was originally designed to annotate short phrases by their associated data practices, the taxonomy and dataset has been adapted for several other contexts and tasks.

While initial inter-coder agreement rates were similar between our questions and the OPP schema, we found that clarity issues in the OPP scheme held steady week-over-week while our scheme's error rate went down. This was true for both the classifications selected by coders, and the text coders marked as relevant to each question. Figure 4 shows the portion of questions changed after each week in our scheme and OPP. After four rounds of iterative adjustment, we removed the OPP annotations from our coding scheme because we felt each change to the OPP annotations made meaningful comparisons between our coding and the original OPP-115 dataset more difficult.

These results demonstrate how ambiguity and noise can come from the coding scheme or the way in which it is presented. While these results caution against uncritical reuse of the OPP taxonomy in new contexts, they do not apply directly to the OPP dataset. OPP was designed to pick out phrases disclosing data practices from short text samples; our scheme is designed to record the ways in which the full text of a privacy policy implicates the rights and responsibilities of firms and users.

We think that these differences in coder convergence between our scheme and OPP help justify our label selection method and iterative approach. There are two potential exogeneities worth addressing here. First, if training and regular meetings were only reason for convergence then we would expect all questions to converge. The OPP questions questions did not, which might suggest that our labels correspond more closely to the terms contained in a privacy policy. Second, if the difference between the two schemes were the result of our initial drafting choices, we would expect to see lower rates of confusion at the outset. Instead, we see similar rates of initial confusion that improve once we detect and correct our drafting issues, further supporting the notion that our labels are better-aligned with policy content.

4 Dataset Contents

The dataset comprises annotations for 162 documents, 88 of which were coded by at least two coders and 74 of which were coded by a single coder. For each document and question, the dataset contains classifications selected by each coder, along with a list of sentences the coder marked as relevant to answering the question and their self-reported confidence ranked on a Likert scale. The dataset includes 64 variables motivated by 11 legal categories.

For a subset of documents and questions, the dataset also includes the amount of time each coder spent answering the question. Coders were all upper-level law students who had completed coursework covering the relevant topics in contract law. Table 1 provides additional descriptive statistics about the dataset.

Except for the CCPA category, correlation between question responses is low, as shown in Figure 5. Since answers to one question are not highly predictive of answers to another, we feel that this set of questions provides reasonable coverage over distinct legal interpretation tasks.

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: Summary statistics on the corpus at time of submission to GenLaw.

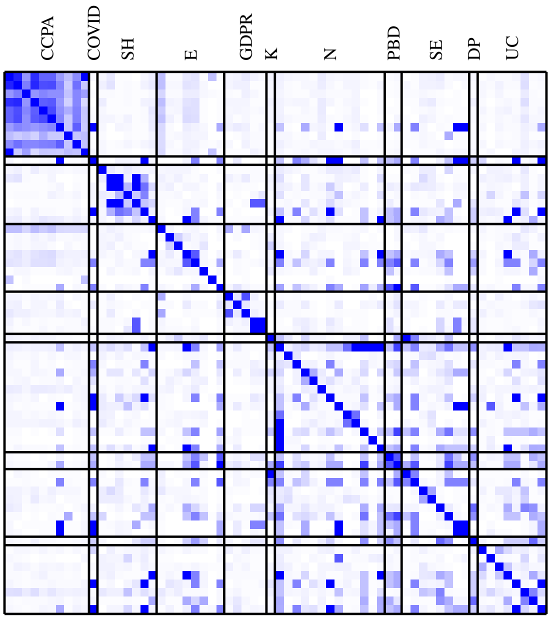


Figure 5: Correlation between question codings.

While achieving high intercoder agreement was explicitly not our goal, the dataset exhibits a moderate to high level of intercoder agreement, with Krippendorff's Alpha ranging from 0.40 to 0.96, averaging at 0.55. Absolute agreement rates across the dataset are 79%, though most disagreement is concentrated in 7 questions. This may suggest that there are a few areas where firms are more likely to use opaque language when describing their terms.

5 Results

Using our dataset, we evaluate models on their ability to perform two tasks. The first task ("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. The second task ("highlight prediction") is an annotation task that targets individual paragraphs: given a question from our dataset and a paragraph from one of the policies in our sample, predict whether a coder marked that paragraph as relevant to answering the question.

	Average BCE Loss		
Category	Claude 3	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 (40k tokens) to perform the task.

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, $\{L_1, L_2, ..., L_n\}$, 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 $\mathbf{y}_i^k$ as $([c_{i1}^k, ..., c_{im}^k])$ normalized to sum to 1, where c_{ij}^k 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_i^k 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, $\frac{1}{m_i}\sum_{j=1}^{m_i}(y_{ij}^k \log(p_{ij}^k)+(1-y_{ij}^k)\log(1-p_{ij}^k))$. We record the average loss by question, category, and across the entire dataset.

Because some policies contain more than 32 thousand tokens, 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. The performance of the major commercial LLMs with sufficiently large context windows appears in table 2. For some of the questions in our dataset, the models performed worse than random guessing, a result we found surprising. The errors appear to be caused by the models incorrectly selecting "not applicable" and "does not disclose" options far too often.

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 f_1. Because highlighting is noisy and heavily skewed, we suspect a certain number of false positives are unavoidable.

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 f_1 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."

6 Future Directions

This project is designed to contribute to the growing body of legal task corpora. We plan to add it to open-source legal benchmarks, such as the LegalBench consolidated corpus.

One of the challenges of analyzing legal documents is how work-intensive it is, how ephemeral some documents are, and how difficult it can be to comparing documents across time, especially if they incorporate changing (external) legal contexts by reference. By releasing these tools and putting greater emphasis on reproducability, we plan to extend this dataset to observe how privacy policies change over time.

Our tools for classifying legal documents were intentionally designed to apply to other legal tasks, or to support future extensions of the question set and dataset. We hope to partner with other legal experts to expand this dataset to cover a broader range of legal questions and documents.

Finally, we have begun investigating the underlying cause of uneven rates of disagreement by question among coders (and, to a lesser extent, similarly uneven response rates between state-of-the-art LLMs). We would like to determine whether firms are intentionally ambiguous to obscure practices or add flexibility, whether a mismatch between technology and law makes certain disclosure difficult or nonsensical, or whether some other factor is at play.

7 Conclusion

We have described our motivation, creation method, and initial analysis of a hand-coded dataset for the interpretation of privacy policies. This new dataset the captures granular multi-class data about 162 privacy policies and their associated documents along 64 dimensions provides a new resource for the development and benchmarking of NLP systems that interpret long legal text. Our coding approach is designed to capture complexities inherent to the task of legal interpretation that are not present in current privacy policy datasets, such as addressing textual ambiguity, indeterminate meaning, interdependent clauses, contractual silence, and the effect of legal defaults. Along with our classification data, we include relevant-text annotations and confidence scores from each labeller. We supplement this dataset with our own coding of the questions where labellers disagree or report low confidence, which may provide additional insight into the textual ambiguities in the underlying policies.

We include the tools we used to produce this dataset, including a hosted online tool that (non-technical) domain experts can use to produce similar classification datasets in other areas of expertise.

Limitations (not part of page limit)

All but one of our population of coders learned consumer contracts (the relevant class for privacy policy interpretation) at the same law school from the same two law professors. They may have adopted some of those professors biases, or approach contract interpretation in similar ways. That overlap may have obscured lingering ambiguities in our coding scheme. It may also have biased them towards understanding a coding scheme designed by one of those professors. We think the latter possibility is fairly remote–privacy policies are a sufficiently esoteric corner of contract law; it receives very little dedicated class time.

As noted above, the "ground truth" meaning of a contract can be probabilistic. Our coders effectively took a noisy sample of each contract with $n = 2$ or $n = 1$. (At time of GenLaw submission, 74 contracts are still $n = 1$, but we RAs ready to fill in those blanks over the summer). We think this is reasonable for two reasons. First, the two-coder approach matches the current state of the art for privacy policy datasets. Second, confidence seems to be a decent predictor of disagreement, which opens mitigation options. We didn't, but potentially could, explore mitigation options.

The noisiness of our measurements also means that our benchmarks in part 5 necessarily contain a (somewhat arbitrary) smoothing factor. We suspect that specific tasks might be better measured using other metrics, and that the smoothing factor could be tuned to reflect confidence.

Likert scores are notoriously messy, meaning our confidence measurements may not contain as much information as we'd ideally like to capture.

Our reported benchmark performance rely on the quality of our prompt design. We have more experience designing prompts for LEGALBERT and GPT-4; our measurement of Claude 3 and other BERT models may be influenced by a prompt that is better suited for GPT. (our prompts are included in our appendices and github repo for reference).

References

Wasi Uddin Ahmad, Jianfeng Chi, Tu Le, Thomas Norton, Yuan Tian, and Kai-Wei Chang. 2021. Intent classification and slot filling for privacy policies.

Wasi Uddin Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. Policyqa: A reading comprehension dataset for privacy policies.

Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021*, pages 2165–2176.

Michael Bommarito II and Daniel Martin Katz. 2022. Gpt takes the bar exam. *arXiv preprint arXiv:2212.14402*.

Duc Bui, Kang G. Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies*, 2021(2):88–110.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english.

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

Jonathan H Choi and Daniel Schwarcz. 2023. Ai assistance in legal analysis: An empirical study. *Available at SSRN 4539836*.

Kevin E Davis and Florencia Marotta-Wurgler. 2019. Contracting for personal data. *NYUL Rev.*, 94:662.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning.

IAB Tech Lab. 2024. IAB Content Taxonomy 3.0.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

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

Florencia Marotta-Wurgler. 2016a. Self-regulation and competition in privacy policies. *The Journal of Legal Studies*, 45(S2):S13–S39.

Florencia Marotta-Wurgler. 2016b. Understanding Privacy Policies: Content, Self-Regulation, and Markets. SSRN Scholarly Paper ID 2736513, Rochester, NY.

Florencia Marotta-Wurgler. 2016c. Understanding privacy policies: Content, self-regulation, and markets. *NYU Law and Economics Research Paper*, (16-18).

Eric Martínez. 2024. Re-evaluating gpt-4's bar exam performance. *Artificial Intelligence and Law*, pages 1–24.

David Mellinkoff. 2004. *The language of the law*. Wipf and Stock Publishers.

Elizabeth Mertz. 2007. *The language of law school: learning to" think like a lawyer"*. Oxford University Press, USA.

Najmeh Mousavi Nejad, Pablo Jabat, Rostislav Nedelchev, Simon Scerri, and Damien Graux. 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*, pages 370–383. Springer.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2020. When does label smoothing help?

Ellen Poplavska, Thomas B Norton, Shormir Wilson, and Norman Sadeh. 2020. From prescription to description: Mapping the gdpr to a privacy policy corpus annotation scheme. In *Legal Knowledge and Information Systems-JURIX 2020: 33rd Annual Conference*.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives.

Joel R Reidenberg, Jaspreet Bhatia, Travis D Breaux, and Thomas B Norton. 2016. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, 45(S2):S163–S190.

Cristiana Santos, Arianna Rossi, Lorena Sanchez Chamorro, Kerstin Bongard-Blanchy, and Ruba Abu-Salma. 2021. Cookie banners, what's the purpose? analyzing cookie banner text through a legal lens. In *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*, pages 187–194.

Atreya Shankar, Andreas Waldis, Christof Bless, Maria Andueza Rodriguez, and Luca Mazzola. 2023. Privacyglue: A benchmark dataset for general language understanding in privacy policies. *Applied Sciences*, 13(6):3701.

Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021. Privacy at scale: Introducing the privaseer corpus of web privacy policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340.

Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, and Noah A. Smith. 2018. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Trans. Web*, 13(1).

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

Appendices

A Question List

Category	Question ID	Question
Contract (K)	K-1	Do the Terms of Use or Terms of Service incorporate the Privacy Policy by reference?
Notice (N)	N-1	’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?’
Notice (N)	N-2	’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? ’
Notice (N)	N-3	’Does the privacy policy state that the company collects or stores biometric information such as facial scans fingerprints facial patterns voice or typing cadence?’
Notice (N)	N-4	’Does the Privacy Policy include a statement noting that that personally identifiable nformation 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.’
Notice (N)	N-5	’Does the privacy policy include a commitmen t 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? ’
Notice (N)	N-6	Are third parties llowed to place advertisements that may track user behavior?
Notice (N)	N-7	Does the privacy policy identifiy third party recipients of shared or sold data?
Notice (N)	N-8	’Does the privacy policy define words such as "affiliates" or "third parties " if it uses them? ’
Sharing (SH)	SH-1	’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?’
Sharing (SH)	SH-2	’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? ’
Sharing (SH)	SH-3	Are third parties bound by the same privacy policy?
Sharing (SH)	SH-4	Does the company perform due diligence to ensure the legitimacy of third parties that may have access to personally identifiable information?
Sharing (SH)	SH-5	’Does the company have a contract with third parties other than processors or service providers establishing how the shared persinally identifiable data can be used?’
Sharing (SH)	SH-6	’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’
Sharing (SH)	SH-7	’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?’
Notice (N)	N-9	’Does the privacy policy include a "Change of Terms " or modifiacation provision that allows the firm to change the privacy policy?’
Notice (N)	N-10	Does the privacy policy require the user/consumer to explicitly assent to any material changes?

Category	Question ID	Question
Notice (N)	N-11	Does the privacy policy states that material changes made to the policy will be retroactive or apply to previous data collection?
User Control (UC)	UC-1	’Does the privacy policy allow the user or consumer to request that incorrect data be either rectified updated or erased?’
User Control (UC)	UC-2	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 ansswer to this question may also be found by exploring the privacy settings of the service.
User Control (UC)	UC-3	Does the privacy policy allow users or consumers to access and correct or update any personally identifiable information collected by the company?
User Control (UC)	UC-4	Does the privacy policy allow the user to request that personally identifiable information be deleted or anonymized?
Notice (N)	N-12	Does the privacy policy summarize the key terms at the top of the policy? Just a table of contents doesn’t count.
User Control (UC)	UC-5	Do the Term of Use include a term explaining the ownership of the user’s or consumer’s personally identifiable information?
Security (SE)	SE-1	Is there a term in the Privacy Policy or Terms of Use guaranteeing data accuracy?
Security (SE)	SE-2	Does the privacy policy specify any reasonable procedures the company many have in place to ensure data accuracy?
Security (SE)	SE-3	Does the privacy policy note whether the firm reserves a right to disclose protected personally identifiable information to comply with law or prevent crime?
Security (SE)	SE-4	Does the firm preserve the right to disclose protected identifiable information to protect its own rights?
Security (SE)	SE-5	Will users or consumers be given notice of any government requests for information about the user?
Security (SE)	SE-6	Does the privacy policy state whether the user will be notified in case there is a data breach?
User Control (UC)	UC-6	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?
Notice (N)	N-13	’Does the privacy policy explain any data procedures if company is sold or otherwise ceases to exist by for example filing for bankruptcy? ’
User Control (UC)	UC-7	’If company is sold or goes bankrupt is the user or consumer given choice as to what happens to their data or personally identifiable information?’
User Control (UC)	UC-8	Does the privacy policy explain what happens to the personally identifialbe infomration of a user who quits the service or closes the account?
Data Practices (DP)	DP-1	Does company have a procedure for safely disposing unused or no longer needed data or personally identifiable information?
Security (SE)	SE-7	’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 o rrequiring periodic audits? ’
Security (SE)	SE-8	’Does the privacy policy identify which means of technological security it employs such as encryption?’
Enforcement (E)	E-1	Does the privacy policy provides means by which user can contact the company with any privacy concerns or complaints? Please select all that apply.
Enforcement (E)	E-2	’Do the privacy policy or the Terms of Use have a forum selection clause? If so which forum? ’
Enforcement (E)	E-3	’Do the privacy policy or Terms of Use have choice of law clause? If so which law? ’
Enforcement (E)	E-4	Do the privacy policy or Terms of Use have an arbitration clause?
Enforcement (E)	E-5	Do the privacy policy or terms of use have a class action waiver?
Enforcement (E)	E-6	Do the privacy policy or terms of use disclaims liability for failure of security measures?

Category	Question ID	Question
Enforcement (E)	E-7	Does the privacy policy provides a link to the Federal Trade Commision's Consumer Complaint Form or does it inlcude t he FTC telephone number?
Enforcement (E)	E-8	'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). '
Privacy By Design (PBD)	PBD-1	Does the privacy policy require periodic compliance review of structural and technological data security measures?
Privacy By Design (PBD)	PBD-2	'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?'
GDPR (GDPR)	GDPR-1	Does the privacy policy states that it complies with GDPR or it includes section on GDPR compliance?
CCPA (CCPA)	CCPA-1	'Doe the privacy policy include a link to the CCPA section as opposed to in the same privacy policy?'
CCPA (CCPA)	CCPA-2	'Does the privacy policy state that the firm's CCPA policy only applies to California residents? For example does it inlcude 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 (CCPA)	CCPA-3	'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 (CCPA)	CCPA-4	'Does the privacy policy directs California Residents to the CCPA section when describing general non-California exclusive data practices?'
CCPA (CCPA)	CCPA-5	Does the privacy policy offer California residents an opportunity to request all information shared with third parties in the last year?
CCPA (CCPA)	CCPA-6	Does the privacy policy offer California residents a direct link via which to contact site and request information?
CCPA (CCPA)	CCPA-7	Does the privacy policy offer data requests by consumers or users explicitly free of charge?
CCPA (CCPA)	CCPA-8	Does the privacy policy list the categories of personal information sold in the past 12 months?
CCPA (CCPA)	CCPA-9	'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.'
GDPR (GDPR)	GDPR-2	Does the privacy policy state that it complies with EU-US Privacy Shield?
GDPR (GDPR)	GDPR-3	Does the privacy policy state that GDPR terms apply only and exclusively to EU residents?

Category	Question ID	Question
CCPA (CCPA)	CCPA-10	’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"?’
GDPR (GDPR)	GDPR-4	’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 (GDPR)	GDPR-5	’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?’
COVID (COVID)	COVID-1	’Does the privacy policy include any terms related to contact tracing health tracking or other terms in relationship to COVID?’

B Average BCE for Holistic Reading Task

Question ID	Claude 3 "haiku"	GPT-4
CCPA-1	0.1221	0.0974
CCPA-2	0.4428	0.0955
CCPA-3	0.2328	0.1810
CCPA-4	0.3474	0.2599
CCPA-5	0.2911	0.2604
CCPA-6	0.3454	0.1875
CCPA-7	0.3050	0.2324
CCPA-8	0.3307	0.3168
CCPA-9	0.2928	0.1722
CCPA-10	0.3041	0.2102
COVID-1	0.0066	0.0066
DP-1	0.0979	0.3332
E-1	0.1993	0.1439
E-2	0.2015	0.2304
E-3	0.0873	0.0771
E-4	0.1424	0.1147
E-5	0.0879	0.2321
E-6	0.2779	0.1487
E-7	0.0164	0.0006
E-8	0.4248	0.4195
GDPR-1	0.1676	0.0813
GDPR-2	0.0398	0.0430
GDPR-3	0.4921	0.2249
GDPR-4	0.3523	0.2689
GDPR-5	0.2108	0.2207
K-1	0.1117	0.0862
N-1	0.1020	0.1016
N-2	0.1322	0.1946
N-3	0.3703	0.3262
N-4	0.4447	0.1618
N-5	0.4528	0.4427
N-6	0.1377	0.1094
N-7	0.3442	0.2391
N-8	0.3565	0.0626
N-9	0.0710	0.0642
N-10	0.4758	0.2301
N-11	0.6175	0.0625
N-12	0.5920	0.1040
N-13	0.5874	0.5340
PBD-1	0.3580	0.5466
PBD-2	0.1504	0.0575
SE-1	0.0740	0.0607
SE-2	0.6783	0.5976
SE-3	0.0938	0.0369
SE-4	0.1960	0.1375
SE-5	0.0407	0.2867
SE-6	0.4634	0.4523
SE-7	0.3876	0.3005
SE-8	0.0922	0.0874
SH-1	0.2494	0.1998
SH-2	0.2139	0.2111
SH-3	0.1823	0.1611
SH-4	0.5571	0.5248
SH-5	0.3440	0.2983
SH-6	0.4556	0.3863
SH-7	0.3277	0.2868
UC-1	0.4287	0.2785
UC-2	0.4593	0.2353
UC-3	0.2130	0.1401
UC-4	0.2137	0.1816
UC-5	0.4042	0.2940
UC-6	0.3517	0.1501
UC-7	0.8190	0.0472

Question ID	Claude 3 "haiku"	GPT-4
UC-8	0.3507	0.3002

C Performance on Highlighting Task

C.1 BERT

Question ID	F1	Recall	Accuracy	Precision
CCPA-1	0.0070	0.0301	0.9585	0.0040
CCPA-10	0.00	0.1984	0.6220	0.0026
CCPA-2	0.0195	0.0314	0.9921	0.0151
CCPA-3	0.0211	0.3240	0.3554	0.0115
CCPA-4	0.0047	0.0035	0.9926	0.0071
CCPA-5	0.0119	0.1278	0.9118	0.0066
CCPA-6	0.0215	0.0752	0.9560	0.0126
CCPA-7	0.0055	0.0095	0.9900	0.0039
CCPA-8	0.0091	0.1386	0.9134	0.0049
CCPA-9	0.0140	0.3024	0.7292	0.0072
COVID-1	0.0011	0.0362	0.3553	0.0005
DP-1	0.0120	0.0090	0.9969	0.0181
E-1	0.0340	0.2729	0.7820	0.0196
E-2	0.0276	0.0470	0.9936	0.0221
E-3	0.0301	0.0790	0.9910	0.0263
E-4	0.0200	0.0561	0.9700	0.0127
E-5	0.0067	0.0168	0.9960	0.0042
E-6	0.0184	0.2299	0.8604	0.0097
E-7	0.0006	0.0270	0.8381	0.0003
E-8	0.0055	0.1224	0.8440	0.0030
GDPR-1	0.0197	0.0691	0.9646	0.0146
GDPR-2	0.0167	0.0267	0.9877	0.0130
GDPR-3	0.0194	0.0169	0.9934	0.0243
GDPR-4	0.0147	0.0147	0.9970	0.0147
GDPR-5	0.0019	0.1975	0.6744	0.0009
K-1	0.0135	0.0873	0.9630	0.0078
N-1	0.0156	0.6252	0.2177	0.0081
N-10	0.01	0.0508	0.9297	0.0070
N-11	0.0066	0.1113	0.8707	0.0035
N-12	0.0080	0.0368	0.9140	0.0047
N-13	0.0041	0.2774	0.4728	0.0021
N-2	0.0106	0.3712	0.6890	0.0055
N-3	0.0168	0.0126	0.9957	0.0252
N-4	0.0315	0.5232	0.6695	0.0170
N-5	0.0177	0.5631	0.4608	0.0092
N-6	0.0314	0.0856	0.9773	0.0272
N-7	0.0344	0.0682	0.9635	0.0259
N-8	0.0097	0.0250	0.9840	0.0097
N-9	0.0145	0.3592	0.8235	0.0079
PBD-1	0.0062	0.0184	0.9810	0.0038
PBD-2	0.0051	0.0060	0.9912	0.0045
SE-1	0.0109	0.0272	0.9920	0.0068
SE-2	0.0093	0.1353	0.9067	0.0051
SE-3	0.0203	0.0331	0.9798	0.0172
SE-4	0.0133	0.4715	0.7461	0.0068
SE-5	0.0117	0.0262	0.9893	0.0085
SE-6	0.0057	0.0267	0.9680	0.0032
SE-7	0.0189	0.1036	0.9531	0.0112
SE-8	0.0075	0.0425	0.9638	0.0044
SH-1	0.0173	0.2206	0.8836	0.0092
SH-2	0.0154	0.0866	0.9513	0.0089
SH-3	0.0307	0.0826	0.9765	0.0203
SH-4	0.0041	0.0704	0.8922	0.0021
SH-5	0.0176	0.1469	0.9413	0.0096
SH-6	0.0195	0.0260	0.9856	0.0212
SH-7	0.0321	0.3043	0.9086	0.0182
UC-1	0.0229	0.0767	0.9597	0.0141
UC-2	0.0119	0.2691	0.7595	0.0062
UC-3	0.0231	0.2021	0.9304	0.0132
UC-4	0.0295	0.0685	0.9809	0.0196
UC-5	0.0606	0.2876	0.9557	0.0416
UC-6	0.0078	0.4496	0.6576	0.0039

Question ID	F1	Recall	Accuracy	Precision
UC-7	0.0198	0.0297	0.9965	0.0148
UC-8	0.0147	0.1140	0.9381	0.0081

C.2 LEGAL-BERT

Question ID	F1	Recall	Accuracy	Precision
CCPA-1	0.0070	0.0301	0.9585	0.0040
CCPA-10	0.0050	0.1984	0.6220	0.0026
CCPA-2	0.0195	0.0314	0.9921	0.0151
CCPA-3	0.0211	0.3240	0.3554	0.0115
CCPA-4	0.0047	0.0035	0.9926	0.0071
CCPA-5	0.0119	0.1278	0.9118	0.0066
CCPA-6	0.0215	0.0752	0.9560	0.0126
CCPA-7	0.0055	0.0095	0.9900	0.0039
CCPA-8	0.0091	0.1386	0.9134	0.0049
CCPA-9	0.0140	0.3024	0.7292	0.0072
COVID-1	0.0011	0.0362	0.3553	0.0005
DP-1	0.0120	0.0090	0.9969	0.0181
E-1	0.0340	0.2729	0.7820	0.0196
E-2	0.0276	0.0470	0.9936	0.0221
E-3	0.0301	0.0790	0.9910	0.0263
E-4	0.0200	0.0561	0.9700	0.0127
E-5	0.0067	0.0168	0.9960	0.0042
E-6	0.0184	0.2299	0.8604	0.0097
E-7	0.0006	0.0270	0.8381	0.0003
E-8	0.0055	0.1224	0.8440	0.0030
GDPR-1	0.0197	0.0691	0.9646	0.0146
GDPR-2	0.0167	0.0267	0.9877	0.0130
GDPR-3	0.0194	0.0169	0.9934	0.0243
GDPR-4	0.0147	0.0147	0.9970	0.0147
GDPR-5	0.0019	0.1975	0.6744	0.0009
K-1	0.0135	0.0873	0.9630	0.0078
N-1	0.0156	0.6252	0.2177	0.0081
N-10	0.0116	0.0508	0.9297	0.0070
N-11	0.0066	0.1113	0.8707	0.0035
N-12	0.0080	0.0368	0.9140	0.0047
N-13	0.0041	0.2774	0.4728	0.0021
N-2	0.0106	0.3712	0.6890	0.0055
N-3	0.0168	0.0126	0.9957	0.0252
N-4	0.0315	0.5232	0.6695	0.0170
N-5	0.0177	0.5631	0.4608	0.0092
N-6	0.0314	0.0856	0.9773	0.0272
N-7	0.0344	0.0682	0.9635	0.0259
N-8	0.0097	0.0250	0.9840	0.0097
N-9	0.0145	0.3592	0.8235	0.0079
PBD-1	0.0062	0.0184	0.9810	0.0038
PBD-2	0.0051	0.0060	0.9912	0.0045
SE-1	0.0109	0.0272	0.9920	0.0068
SE-2	0.0093	0.1353	0.9067	0.0051
SE-3	0.0203	0.0331	0.9798	0.0172
SE-4	0.0133	0.4715	0.7461	0.0068
SE-5	0.0117	0.0262	0.9893	0.0085
SE-6	0.0057	0.0267	0.9680	0.0032
SE-7	0.0189	0.1036	0.9531	0.0112
SE-8	0.0075	0.0425	0.9638	0.0044
SH-1	0.0173	0.2206	0.8836	0.0092
SH-2	0.0154	0.0866	0.9513	0.0089
SH-3	0.0307	0.0826	0.9765	0.0203
SH-4	0.0041	0.0704	0.8922	0.0021
SH-5	0.0176	0.1469	0.9413	0.0096
SH-6	0.0195	0.0260	0.9856	0.0212
SH-7	0.0321	0.3043	0.9086	0.0182
UC-1	0.0229	0.0767	0.9597	0.0141
UC-2	0.0119	0.2691	0.7595	0.0062
UC-3	0.0231	0.2021	0.9304	0.0132

Question ID	F1	Recall	Accuracy	Precision
UC-4	0.0295	0.0685	0.9809	0.0196
UC-5	0.0606	0.2876	0.9557	0.0416
UC-6	0.0078	0.4496	0.6576	0.0039
UC-7	0.0198	0.0297	0.9965	0.0148
UC-8	0.0147	0.1140	0.9381	0.0081