Evaluating Privacy Policies under Modern Privacy Laws At Scale: An LLM-Based Automated Approach

Qinge Xie Karthik Ramakrishnan Frank Li
Georgia Institute of Technology

Abstract

Website privacy policies detail an online service's information practices, including how they handle user data and rights. For many sites, these disclosures are now necessitated by a growing set of privacy regulations, such as GDPR and multiple US state laws, offering visibility into privacy practices that are often not publicly observable. Motivated by this visibility, prior work has explored techniques for automated analysis of privacy policies and characterized specific aspects of real-world policies on a larger scale. However, existing approaches are constrained in the privacy practices they evaluate, as they rely upon rule-based methods or supervised classifiers, and many predate the prominent privacy laws now enacted that drastically shape privacy disclosures. Thus, we lack a comprehensive understanding of modern website privacy practices disclosed through privacy policies.

In this work, we seek to close this gap by providing a systematic and comprehensive evaluation of website privacy policies at scale. We first systematize the privacy practices discussed by 10 notable privacy regulations currently in effect in the European Union and the US, identifying 34 distinct clauses on privacy practices across 4 overarching themes. We then develop and evaluate an LLM-based approach for assessing these clauses in privacy policies, providing a more accurate, comprehensive, and flexible analysis compared to prior techniques. Finally, we collect privacy policies from over 100K websites, and apply our LLM method to a subset of sites to investigate in-depth the privacy practices of websites today. Ultimately, our work supports broader investigations into web privacy practices moving forward.

1 Introduction

User privacy on the web has become a top-level public concern, as evidenced by the increasing adoption of regional privacy regulations including the European Union's General Data Protection Regulation (GDPR) and a growing number of US state laws (such as in California, Texas, and Florida). As a component of these privacy laws, many online services are required to disclose their privacy practices through privacy policy documents, including discussing how they collect, use, and share user information, as well as the rights users can exercise in managing their data. Because many of these practices are not externally observable, privacy policies offer a rare window into the privacy behaviors of online services.

Capitalizing on this visibility is challenging though, as privacy policies are lengthy, complex, free-form documents. Prior work has explored techniques for automated privacy policy analysis. However, these approaches are either unsupervised and rule-based [10, 17, 20, 68, 69], or they involve trained/fine-tuned classifiers [7, 34, 55, 75], and their capabilities are constrained to analyzing a limited set of pre-defined privacy clauses in policies. Furthermore, many works predate modern privacy regulations [67, 68], or were evaluated on datasets predating modern regulations [32, 34, 64], and do not fully encapsulate the privacy practices discussed in current laws. Although these methods have been applied to evaluate specific characteristics of real-world policies on a larger scale [38, 50, 64], due to their limitations, we ultimately still lack a comprehensive and modern understanding of web privacy practices as disclosed through these policies.

In this work, we seek to fill this gap through a systematic, comprehensive, and at-scale evaluation of website privacy policies, characterizing what privacy practices are disclosed and assessing in-depth privacy behaviors. To ground our evaluation, we first systematize the privacy clauses discussed by 10 notable privacy laws already in effect, including the EU's GDPR and 9 US state laws. Prior work involving privacy clause systematization focused on narrower sets of clauses, often limited to a single law (typically GDPR), or selected subsets from GDPR and CCPA. Across the ten laws, we identify 34 distinct clauses on privacy practices, while existing human-annotated datasets cover only half of them. We group these clauses into 4 overarching themes (personal information practices, explanation of user rights, methods for exercising rights, and disclosure of specific types of information).

Guided by this systematization, we then develop an ap-

proach using Large Language Models (LLMs) for assessing these clauses in privacy policies. LLMs offer state-of-the-art performance in text comprehension, reasoning, and question-answering. Our systematic evaluation demonstrates that our LLM method provides accurate, comprehensive, and flexible analysis compared to prior techniques. For all tasks across the 34 privacy clauses, our LLM approach exhibits an average F1-score of 0.94 (with no task below 0.84), whereas the F1-score for existing methods is much lower (mostly below 0.8). Beyond improved performance, our analysis pipeline uses open-source models and thus serves as a platform for future privacy policy studies.

Finally, we collect privacy policies from the Google CrUX top 100K sites as well as domains for the top 1K US and top 500 EU companies (according to Fortune 2024). We apply our LLM method to a subset of sites, particularly those with English privacy policies, to investigate in-depth the privacy practices of thousands of websites today. We characterize what privacy clauses are discussed in the policies, as well as some of the detailed privacy practices disclosed. For example, we identify that while policies broadly discuss clauses related to personal information practices, fewer discuss other categories, such as user rights and methods for exercising them. We also observe broader coverage of clauses by higher-ranked domains, as well as the influence of regional privacy regulations. Furthermore, we quantify the types of personal information that sites collect, their primary purposes, and the third-party entities with whom the data is shared (with Google entities being particularly prominent).

Ultimately, our study provides a more comprehensive and modern understanding of website privacy practices via their privacy policies and provides a foundation for broader investigations into web privacy moving forward. In summary, our primary contributions include:

- Systematization of the privacy policy clauses across 10 modern privacy laws (GDPR and 9 US state laws).
- An LLM approach for accurately and comprehensively evaluating privacy clauses in website privacy policies.
- A larger-scale measurement of real-world privacy policies on thousands of websites.
- Open-sourced the LLM-PP2025 (LLM–Privacy Policy 2025) dataset and LLM-based analysis pipeline at https://github.com/BEESLab/LLM-PP2025.

2 Related Work

Here, we discuss related work on analyzing privacy policies.

Automated Methods for Privacy Policy Analysis. Several prior studies have explored automated analysis methods for privacy policies. Some of these approaches are unsupervised and rule-based [10, 17, 20, 69, 74], often involving the construction of ontologies to define subsumptive relationships between terms in privacy policies and employ NLP techniques. For example, PolicyLint [10] extracts data types and entities from privacy policies for contradiction analysis. PurPliance [17] analyzes the predicate-argument structure of policy sentences and classifies the extracted purpose clauses into a taxonomy of data purposes. Building on rule-based methods, PoliGraph [20] further leverages knowledge graphs to capture context across different parts of privacy policies.

Another category of approaches primarily utilizes annotated privacy policy datasets (e.g., OPP-115 [67], APP-350 [76], and PPGDPR [46]) to train machine learning classifiers for privacy policy analysis [23], such as using Support Vector Machines [76], Logistic Regression [14, 58], Convolutional Neural Networks [34], BERT-based models [55, 60, 64], and XLNet models [7]. Both the rule-based approaches and the machine learning classifiers are constrained to analyzing a limited set of pre-defined privacy clauses in policies.

With the emergence of LLMs, researchers have applied them to legal reasoning tasks. Guha et al. [32] evaluated the performance of 20 LLMs on various legal tasks beyond just privacy policies. Chen et al. [18] developed a tool that automatically detects privacy policy snippets and potential risks related to user-input sensitive data. Rodriguez et al. [57] assessed ChatGPT and the Llama 2 model on identifying and categorizing several data practices in privacy policies. PolicyGPT [61] and Goknil et al. [29] evaluated LLM performance on the OPP-115 and PPGDPR datasets, exploring different prompt engineering configurations[1]. However, these studies mainly focus on evaluating LLM performance on existing annotated datasets. Unlike our work, they do not account for the privacy clauses from modern privacy laws, and they do not apply their methods for large-scale measurements.

Large-Scale Measurements of Privacy Policies. Prior privacy policy measurements [8, 15, 53, 59, 68] typically focus on limited characteristics of policies. For example, Amos et al. [8] analyzed how privacy policy properties (e.g., length) change over time. Bui et al. [16] analyzed the inconsistencies between web trackers' data practices and the opt-out statements in their privacy policies. PolicyChecker [68] used a rule-based approach to assess the GDPR compliance of mobile app privacy policies. Wagner [64] conducted a large-scale longitudinal study of the contents of privacy policies using the categories and attributes from the OPP-115 dataset. Our work seeks to provide a more comprehensive, accurate, and modern understanding of the privacy practices disclosed in real-world privacy policies.

Comparing Privacy Policies with Actual Data Collection. Another type of measurement work [11, 26, 37, 63] focuses on comparing the data collection statements in a business's privacy policy with its actual data collection behaviors observed in network traffic. For example, OVRSEEN [63] analyzed data flows found in network traffic from Oculus VR

[1] We note that PolicyGPT [61] and [29] have not been peer-reviewed yet. In fact, PolicyGPT [61] has transparency/reproducibility issues raised by [29]. Thus, we do not discuss the performance of these works in Section 4.3.

No.	Privacy Law	Eff. Date	Abbr.
1	General Data Protection Regulation (GDPR)	05/25/2018	GDPR
2	California Consumer Privacy Act (CCPA)	01/01/2020	CA
3	Virginia Consumer Data Protection Act (VCDPA)	01/01/2023	VA
4	Colorado Privacy Act (CPA)	07/01/2023	CO
5	Connecticut Data Privacy Act (CDPA)	07/01/2023	CT
6	Utah Consumer Privacy Act (UCPA)	12/31/2023	UT
7	Florida Digital Bill of Rights (FDBR)[2]	07/01/2024	FL
8	Oregon Consumer Privacy Act (OCPA)	07/01/2024	OR
9	Texas Data Privacy and Security Act (TDPSA)	07/01/2024	TX
10	Montana Consumer Data Privacy Act (MCDPA)	10/01/2024	MT

Table 1: Privacy laws analyzed and their effective dates.

apps and compared them with the data collection statements in the privacy policies. Iqbal et al. [37] analyzed the consistency between the data collection practices of Alexa skill vendors (observed in network traffic) and the statements made in their privacy policies. Our work does not examine actual data collection, but it can support future work in this direction by providing a deeper understanding of privacy policy statements.

3 Systematization of Privacy Policy Clauses

To ground our evaluation of privacy policies, here we first systematize the privacy policy clauses that are discussed across modern privacy laws. Specifically, we investigate the 10 privacy laws in effect for the European Union (GDPR) and 9 different states within the US [21], which all require businesses to provide privacy policies for data transparency. Table 1 lists these laws and their effective dates.

Prior Systematization. Prior work involving systematization focused on narrower sets of privacy clauses (particularly from one law). Since its enactment in 2018, the GDPR's privacy policy requirements have been extensively examined by the research community from multiple perspectives [72], such as GDPR's impact on privacy policies [8, 22, 45], GDPR-completeness violations [31, 33, 68], and GDPR compliance [25, 56]. Additionally, GDPR-specific human-annotated datasets have been created (e.g., PPGDPR [46]), enabling researchers to evaluate their automated approaches to privacy policy analysis, which involved first systematizing GDPR requirements and then annotating privacy segments accordingly. CCPA has been studied to a lesser extent. Prior work involving systematization and comparison with GDPR includes more narrow sets of privacy clauses [36, 47]. The C3PA dataset [49], created in 2024, was the first open CCPA-aware dataset of human-annotated privacy policies. The growing number of US state laws further requires a modern, comprehensive systematization of all these privacy laws. Meanwhile, while California crafted its own privacy law, other states initially based their laws on a version of the yet-to-pass Washington Privacy Act (WPA), which was introduced in 2019 [51]. As a result, these WPA-inspired laws share similar language across many clauses, while CCPA remains an outlier in several respects, as will be discussed in this section. Finally, as will be discussed in Section 4, existing annotated datasets cover only half of the clauses we study in this work. Our systematization is more expansive and modern, and can guide future privacy studies.

Systematization Process. Each law considered contains specific sections outlining privacy policy requirements. A team of three web privacy experts comprehensively reviewed all provisions, as well as the full texts of each law, then met to discuss and converge on the systematization (this manual systematization process is similar to prior work [46]). Before convergence, the initial inter-rater agreement (Fleiss's Kappa) was 0.882. Across the 10 laws, we identified 34 distinct clauses specifying content requirements for privacy policies. As shown in Table 2, we group these clauses into four thematic categories: 1) A comprehensive description of a business's *personal information practices* (P1-7), 2) Information on consumer *rights* with their personal information (R1-10), 3) Instructions for *exercising* those rights (E1-8), and 4) *Disclosures* about specific types of information (D1-9). Table 2 also lists which laws specify each clause. For reference, we list the provision number corresponding to each clause in our GitHub repository.

We note that not all clauses apply to all privacy policies (e.g., GDPR targets EU-based companies or companies with EU consumers). In this work, we do not seek to audit whether privacy policies fully adhere to applicable regulations (a promising direction for future work, which would require identifying which laws apply to each online service). Instead, we aim here to distill the clauses that may be included in privacy policies based on enacted privacy laws, guiding our subsequent characterization of privacy policy content.

Below, we describe briefly the clauses identified across the four categories. *We note that our systematization reflects the laws as of January 2025 and remains accurate at the time of writing. The privacy laws studied may be amended in the future and, as a result, affect the systematized clauses.*

Personal Information Practices. All privacy laws require that privacy policies describe personal information practices. We observed the following clauses across the 10 privacy laws.

- **P1-2**: All privacy laws require that a privacy policy discloses the categories of a consumer's personal information collected or processed by the business (P1). GDPR mandates P1 when personal data is collected indirectly (e.g., via third parties) but not directly from the consumer.
  Additionally, all US state laws require businesses to provide the categories of personal data they share, sell, or disclose to third parties (P2), although GDPR does not. Note that laws differ in whether they distinguish sharing, selling, and disclosing data. Given the inconsistencies and overlap in definitions, we group all three actions together.

[2]The FDBR took effect on July 1, 2024. It also includes a provision aimed at curbing government influence online, which took effect on July 1, 2023.

ID	Clause	GDPR	CA	VA	CO	CT	UT	FL	OR	TX	MT
	Personal Information Practices										
P1	Categories of PI Collected	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
P2	Categories of PI Shared/Sold/Disclosed		✓	✓	✓	✓	✓	✓	✓	✓	✓
P3	Categories of PI Sources	✓	✓								
P4	Purpose for Collecting PI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
P5	Purpose for Selling/Sharing/Disclosing PI		✓								
P6	Categories of Third-Party Recipients of PI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
P7	PI Retention Period	✓									
	Explanation of Consumer **R**ights										
R1	Right to Know (Access, Confirm)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
R2	Right to Data Portability	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
R3	Right to Delete (Erase)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
R4	Right to Correct (Rectify)	✓	✓	✓	✓	✓		✓	✓	✓	✓
R5	Right to Opt-Out (Ads/Sales/Profiling)		✓	✓	✓	✓	✓	✓	✓	✓	✓
R6	Right to Object (General Processing)	✓									
R7	Right to Limit (Restrict Processing)	✓	✓								
R8	Right to Lodge a Complaint	✓									
R9	Right to Withdraw Consent	✓									
R10	Right to Non-Discrimination		✓								
	Methods for **E**xercising Consumer Rights										
E1	Methods to Submit A Request		✓	✓	✓	✓	✓	✓	✓	✓	✓
E2	Process for Consumer to Appeal Controller's Decision			✓	✓	✓		✓	✓	✓	✓
E3	Contact Information	✓	✓		✓	✓			✓		✓
E4	Process for Verifying Consumer Requests		✓								
E5	Instructions for Authorized Agent Requests		✓								
E6	Opt-In Processes for Sale/Sharing of PI of Under-16s		✓								
E7	Explanation of Processing An Opt-out Preference Signal		✓								
E8	Implementation of Frictionless Opt-out Signals		✓								
	Required **D**isclosures of Specific Information										
D1	Notice of Sale of Sensitive PI							✓		✓	
D2	Notice of Sale of Biometric PI							✓		✓	
D3	Statement on Use/Disclosure of Sensitive PI		✓								
D4	Statement on Knowingly Selling/Sharing PI of Under-16s		✓								
D5	Last Updated Date		✓								
D6	Metrics Report from Large PI Collectors		✓								
D7	Third Country Transfer	✓									
D8	Automated Decision-Making	✓									
D9	Statutory/Contractual PI Requirements	✓									

Table 2: Privacy policy clauses as specified across GDPR and 9 US state laws, organized into four thematic categories, along with the privacy laws that mandate these clauses.

- **P3**: GDPR and CCPA require a privacy policy to include information on the categories of sources from which personal information is collected, such as directly from consumers, advertising networks, data brokers, etc.
- **P4-5**: All privacy laws require that the purposes of collecting and processing consumers' personal information be disclosed in a privacy policy (P4). The CCPA further mandates that businesses disclose the purposes for selling, sharing, and disclosing personal information (P5)[3].
- **P6**: All privacy laws mandate information on the categories of third-party entities with whom a business shares, sells, or discloses personal information.
- **P7**: GDPR requires that privacy policies provide a retention period for collected personal information, or the criteria used to determine that period. (CCPA requires retention information to be disclosed in the Notice at Collection, which is not necessarily included in the privacy policy.)

Explanation of Consumer Rights. All privacy laws require privacy policies to include information about the rights consumers have regarding their personal information.

- **R1-5**: All privacy laws entitle consumers to the right to know what personal information a business has collected about them (R1), to receive a copy of their processed personal information in a portable format (R2), to request

[3]GDPR also requires a privacy policy to provide the legitimate interests for data processing. However, legitimate interest is a flexible lawful basis for processing [52] and is the most ambiguous clause in GDPR as it allows for broad interpretations of different processing purposes [42]. Thus, we choose not to separately consider clauses about legitimate interests in this study.

the deletion of their personal information (R3). Except for Utah's privacy law, all others entitle consumers the right to have inaccuracies in their information corrected (R4). Except for GDPR, all other US laws entitle consumers the right to opt out of businesses processing their personal information specifically for targeted advertising, sale of personal data, and profiling (R5).

- **R6**: GDPR entitles consumers to object to/opt-out of the general processing of personal data.
- **R7**: GDPR and CCPA both provide consumers with the right to limit the processing of personal information beyond specific conditions (outlined in each law).
- **R8-9**: GDPR grants consumers the right to lodge a complaint with a supervisory authority (R8), and the right to withdraw consent for processing their personal information at any time (R9).
- **R10**: CCPA gives consumers the right not to receive discriminatory treatment by the business for the exercise of privacy rights.

Methods for Exercising Consumer Rights. All privacy laws further require that a privacy policy provide some information on how to exercise these consumer privacy rights.

- **E1**: All US privacy laws require businesses to describe the methods by which consumers can submit requests in their privacy policy. While GDPR does not explicitly require E1, it has comprehensive regulations on how businesses are to respond to consumer requests.
- **E2**: Most US state laws (except for California's and Utah's) require information on how consumers may appeal a business's decision on their requests.
- **E3**: As required by GDPR and 5 US states, a business's contact information must be included in a privacy policy to allow consumers to exercise their rights and ask questions. Three US states specify that the contact method should be online, such as an email address.
- **E4-8**: California's CCPA includes more detailed requirements on describing the exercise of consumer rights[4], including a description of the process for verifying consumer requests (E4), and instructions on how an authorized agent can make a request on a consumer's behalf (E5). Furthermore, if the business has actual knowledge that it sells the personal information of consumers under 16 years of age, a description of the processes for opting into the sale or sharing of this information should be provided (E6). Another notable aspect is the handling of opt-out signals. CCPA requires an explanation of how an opt-out preference signal will be processed for the consumer (E7), and if the business processes frictionless opt-out preference signals, how consumers can implement the signals (E8).

Specific Disclosures Finally, we identify nine clauses on requirements for disclosing specific information[5].

- **D1-2**: Florida and Texas require businesses to disclose in their privacy policy if they sell *sensitive* personal data (as defined by the laws) (D1) or biometric data (D2).
- **D3**: CCPA requires businesses to state whether they use or disclose *sensitive* personal information (as defined by CCPA) for purposes other than those specified in its Section 7027(m) (e.g., providing expected goods or services).
- **D4**: CCPA requires businesses to state whether they have actual knowledge of selling or sharing personal information of consumers under 16 years old.
- **D5**: CCPA requires that a privacy policy include the date it was last updated.
- **D6**: CCPA requires businesses collecting 10M+ consumers' information annually to include a metrics report or link in their privacy policies, detailing consumer privacy requests received and average response time over the past year.
- **D7**: GDPR requires businesses to disclose whether they intend to transfer personal data to a third country or international organization.
- **D8**: GDPR requires businesses to inform consumers about the existence of automated decision-making, including profiling, and its use of personal data.
- **D9**: GDPR requires businesses to inform consumers whether the provision of personal information is mandatory (e.g., statutory or contractual), and the consequences of not providing such information.

4 LLM-Based Privacy Policy Analysis

In this section, we introduce our LLM-based framework that affords an automated assessment of privacy policies, evaluating the presence of each clause identified in Section 3. LLMs have demonstrated significant advancements across various tasks, including legal reasoning [32, 41, 71]. Multiple LLMs have been developed, such as OpenAI's GPT, Meta's Llama [62], and Google's Gemini [12]. These models possess extensive world knowledge, strong problem-solving capabilities, sophisticated reasoning skills, and a robust capacity for following instructions.

In this work, we use the Llama-3.1-70B-Instruct [6] as the foundational model for interpreting privacy policies. The Instruct version is a fine-tuned version of its text-only model for assistant-like chat. While Llama-3.1-405B [5] is currently the largest and most capable openly available foundation model, its substantially higher hardware requirements make it impractical for most researchers to run. The 70B model delivers comparable performance to the 405B model across different tasks [4] while being more resource-efficient and ensuring that our framework remains accessible and reproducible. Besides, using an open-source model also supports future fine-tuning

[4] CCPA includes two other clauses requiring businesses to provide a notice of the Right to Opt-out of Data Sale/Sharing and a notice of Right to Limit. These two clauses largely overlap with R5 and R7 so we do not consider these clauses separately.

[5] We classify some clauses as Disclosures instead of Personal Information Practices if they are about specific information types, rather than general personal information processing.

needs, and allows us to run inference on our own infrastructure, enabling efficient large-scale measurement.

We design and evaluate the performance of two types of LLM tasks for privacy policy assessment:

1) **Coverage Tasks**: A coverage task assesses whether a privacy policy segment *covers* the required content of a specific clause. We design prompts to guide the LLM to: 1) perform binary classification, and 2) extract texts from the privacy policy segments that are relevant to the specific clause. The prompt template is shown in Figure 4 in Appendix B.
2) **Inspection Tasks**: Building on the coverage tasks, we conduct assessments of personal information practices in privacy policies to gain a deeper understanding of how current businesses disclose the use of consumers' personal information. Specifically, we *inspect* the categories of personal information the business collects (P1) and shares (P2), the purposes (P4, P5), and the third-party recipients (P6). We design prompts to guide the LLM to perform multi-class classification on categories of personal information, purposes, and third-party recipients (discuss in Section 4.2). All prompts for the inspection tasks are available in our GitHub repository.

4.1 Evaluation Dataset for Coverage Tasks

To evaluate our LLM prompts for coverage tasks, we require labeled data on privacy policy text. Our systematized clauses (Section 3) are significantly broader than those considered in prior work [46, 49, 67], and as a result, there is not one complete dataset whose labels cover all clauses. Instead, we use three human-annotated privacy policy datasets created by prior work that together encapsulate about half of our clauses, and we manually label data for the remaining clauses (as shown in Table 3). Here, we describe these existing datasets and how we construct the evaluation dataset.

Existing Datasets. We use three existing datasets:

1) **OPP-115**: The OPP-115 dataset [67] provides annotations for 115 privacy policies. Despite being created before GDPR (the earliest of our 10 laws), it remains the most commonly used dataset for developing automated privacy policy analysis (e.g., Polisis [34]). In the dataset, each data practice in a policy text segment is categorized into one of ten main categories, with further category-specific attributes. We directly map one main category to clause P7, and we map P1-2 and P4-6 to attributes specific to other categories. For example, we map P4 to the "Purpose" attribute in the main category "First Party Collection/Use". We detail the label mapping process in Appendix A.
2) **PPGDPR**: The PPGDPR dataset [46] includes labeled sentences from 304 English mobile app privacy policies, based on GDPR §Art.13 regarding personal information collection. We identify eight labels aligned with clauses P7, R1-2, R6-8, and E3. PPGDPR merges the right to correct (R3) and the right to delete (R4) into a single label, whereas we assess these two clauses separately; thus, we do not use PPGDPR for these two clauses.
3) **C3PA**: The C3PA dataset [49] annotated text segments from 411 organizations' privacy policies with CCPA-specific disclosures. We map eight of its labels to clauses R1, R3-5, R7, R10, E1 and D5.

Evaluation Dataset Construction. Segments/sentences in all three datasets were labeled by three annotators. PPGDPR only provides aggregated labels, while OPP-115 and C3PA provide each annotator's label. To determine a final label in OPP-115 and C3PA, we apply the majority rule, as recommended by OPP-115 [67].

To evaluate task performance, we require positive and negative samples of each clause. For clauses mapped to labels in existing datasets, we select all policy text segments with those labels as positive samples (i.e., clause content is in the privacy policy). We then select an equal number of negative samples by randomly sampling text segments in each dataset without the labels. For clauses not covered by the prior datasets, we continuously randomly sampled and labeled segments until 50 positive and 50 negative samples were found from our collected real-world privacy policy dataset (see Section 5).

4.2 Evaluation Dataset for Inspection Tasks

For the inspection tasks, each privacy policy may use different terminology for describing personal information practices (we specifically focus on P1-2 and P4-6[6]). To avoid analyzing an unbounded number of practice descriptions, we frame the inspection tasks as multi-class classification and define classes of practices (e.g., for P1, classes of PI collected) in LLM prompts. This section details how we define the classes and construct the evaluation dataset.

Class Definitions. As discussed in Section 4.1, OPP-115 contains fine-grained attribute labels that directly map to the personal information practice clauses investigated in our inspection tasks. Thus, we initialize our class definitions with OPP-115 labels. However, as OPP-115 predates modern privacy laws, we observe further definitions, especially in California's CCPA [1]. Thus, we refine OPP-115 classes with CCPA definitions, providing a more comprehensive and updated set of classes. Based on CCPA definitions, we make the following changes to OPP-115 classes:

- For personal information types (P1, P2), we introduce four new classes (e.g., biometric data), and refine the definition of "identifier" using CCPA's expansive definition (whereas OPP-115 considers this class a catch-all for user identifiers not accounted for by other classes).
- For data collection and disclosure purposes (P4, P5), we merge some OPP-115 classes based on CCPA definitions

[6]We omit P3 as the definition of PI sources is vague in existing laws, and P7 as privacy policies can specify either a specific period or a criterion for determining it, and the later widely varies in description.

(e.g., defining advertising and marketing as a single class).
- For third-party recipients, OPP-115 defines four classes besides unnamed and unspecified entities: named entities, affiliates, other users, and the public. Guided by CCPA, we separate OPP-115's "named entities" class into "explicitly named entities" (e.g., Facebook) and "categorized entities" (e.g., data brokers), which merge in "affiliates". We also merge together "other users" and "the public", as OPP-115 segments typically have both labels (e.g., user comments visible to other users and the public).

Evaluation Dataset Construction. We begin with the text segments identified in the coverage task as containing the relevant clauses (P1-2, P4-6). For classes defined already in OPP-115, we select the relevant OPP-115 labels as positive samples and randomly sample an equal number of segments with other labels as negative samples. For new classes not defined in OPP-115, we continuously randomly sampled and label segments until 30 positive and 30 negative samples were found from our collected privacy policy dataset (Section 5). For some classes, we observe few positive samples in OPP-115, so we augment our samples in a similar manual fashion.

4.3 LLM Prompt Engineering and Evaluation

Here, we describe our LLM prompt design and configuration and present our evaluation results for each assessment task.

LLM Prompt Design. LLMs require a prompt instructing them on how to perform a task. Similar to LLMs applied for other legal reasoning tasks [32], we design prompts for each task by manually writing instructions and drawing on the original legal text in privacy laws. As discussed in Section 3, the terms used for the same clause across different privacy laws may vary. Thus, we also incorporate language from different privacy laws to create comprehensive task instructions.

We use zero-shot learning (where prompts do not include examples). While we evaluated few-shot learning and chain-of-thought [40,65] prompting, common LLM prompting techniques, we observed limited improvements or worse performance on some tasks, as detailed in Table 11 in Appendix C.

Input Segmentation. While the privacy policies in the ground truth dataset are manually segmented, online privacy policies are typically complete documents that require segmentation. While Llama 3.1 expands the input context length to 128K tokens, prior research has shown that longer input lengths can lead to performance drops [43]. We also observe in our experiments that inputting the full privacy policy document often leads to false negatives. Thus, we segment privacy policies before inputting to the LLM, and set a 1K token limit for each segment, as performance is only slightly impacted [43], while maintaining the context of the privacy policy. As will be discussed in Section 5, we use the Markdown format of privacy policies as LLM inputs to maintain the HTML table context and the overall structure of a privacy policy. We hierarchically segment the policy by Markdown headings, ensuring each segment remains within 1K tokens. If the lowest heading level is reached and tokens still exceed 1K, we further split the segments by line breaks.

Cl.	Dataset	#Samples	Baseline (F1)	Our's (F1)
P1	OPP-115	1632	N/A	0.85
P2	OPP-115	732	N/A	0.90
P3	Ours	100	N/A	0.84
P4	OPP-115	1820	N/A	0.91
P5	OPP-115	1336	N/A	0.93
P6	OPP-115	848	N/A	0.93
P7	OPP-115/PPGDPR	96/896	0.71/0.82	0.97/0.96
R1	PPGDPR/C3PA	230/2236	0.63/0.57|0.40	0.91/0.94
R2	PPGDPR	334	0.82	0.96
R3	C3PA	1162	0.74	0.97
R4	C3PA	646	0.66	0.98
R5	C3PA	1226	0.69	0.92
R6	PPGDPR	490	0.71	0.88
R7	PPGDPR/C3PA	254/282	0.80/0.73	0.93/0.95
R8	PPGDPR	290	0.84	0.95
R9	Ours	100	N/A	0.91
R10	C3PA	756	0.91	0.99
E1	C3PA	2432	0.83	0.95
E2	Ours	100	N/A	0.95
E3	PPGDPR	1442	0.79	0.84
E4	Ours	100	N/A	0.98
E5	Ours	100	N/A	1.00
E6	Ours	100	N/A	0.95
E7	Ours	100	N/A	0.88
E8	Ours	100	N/A	0.99
D1	Ours	57	N/A	1.00
D2	Ours	51	N/A	1.00
D3	Ours	100	N/A	0.93
D4	Ours	100	N/A	0.98
D5	C3PA	194	0.98	0.99
D6	Ours	100	N/A	1.00
D7	Ours	100	N/A	1.00
D8	Ours	100	N/A	0.94
D9	Ours	100	N/A	0.93

Table 3: F1-scores from evaluating our method's performance across coverage tasks (compared to existing baselines).

LLM Configuration. We deploy the Llama-3.1-70B model using Hugging Face Transformers in bfloat16 precision (i.e., without quantization) on a cluster with 48 NVIDIA H200 GPUs. Each task runs on a node with two H200 GPUs. We set the max token length to 2048 and disable sampling (`do_sample`=false) to obtain deterministic results.

Coverage Task Evaluation. Table 3 lists our LLM's performance (F1-score) on coverage tasks for each clause using zero-shot learning, along with the evaluation dataset used and the total number of test samples (balanced between positive and negative samples[7]). We note that there are other prior works that have evaluated similar datasets like OPP-115, but their evaluations entail largely different tasks/goals,

[7]For D1 and D2, we could only find a small number of positive samples (7 and 1, respectively). Thus, we use these positive samples with 50 negative samples (as done for other tasks/clauses).

which are not equivalent to our coverage tasks. For example, LegalBench [32] evaluated LLM performance on the OPP-115 dataset for the binary classification task of determining whether a segment belongs to one of OPP-115's main categories, which do not directly align with legal clauses (except P7, as discussed in Section 4.1). Thus, as a baseline comparison, if prior work analyzed an *equivalent task*, we also list the F1-scores reported. We only identified comparable evaluations for 12 tasks [32, 46, 49]. For some clauses (e.g., P7), multiple datasets had relevant labels, so we separately listed our (or baseline) performance on each dataset. For R1 and the C3PA dataset specifically, C3PA contains two relevant labels, and we present our task evaluation on each label separately.

Across all tasks/clauses, our LLM method achieved high F1-scores (at least 0.84, but mostly above 0.9), with an average F1-score of 0.94 across tasks. For each task with a baseline comparison, our performance substantially exceeds that reported in prior work [32, 46, 49] (where the average baseline F1-score across these tasks is 0.74). Thus, our approach provides significantly more accurate and comprehensive coverage of modern privacy policy clauses in privacy policies.

While non-equivalent tasks from prior work are not directly comparable, we discuss them to illustrate the relative performance of our approach, given the lack of equivalent baselines. The highest F1-score reported by LegalBench [32] was 0.806, on the First Party Collection/Use category in OPP-115 using GPT-3.5, which relates to clauses P1 and P4. Our approach achieved scores of 0.85 and 0.91, respectively. For the Third Party Sharing/Collection category, LegalBench achieved an F1-score of 0.801 using GPT-4, corresponding to clauses P2, P5, and P6. Our scores were 0.90, 0.93, and 0.93, respectively.

Finally, we do observe worse performance on some tasks compared to others (e.g., P3), which upon manual inspection, appears primarily due to more complexity, diversity, or ambiguity in privacy policy languages that satisfy a clause.

Inspection Task Evaluation. For the inspection tasks, we list our LLM approach's performance doing multi-class classification of personal information types (P1, P2) and purposes (P4, P5) in Table 4 and Table 5, respectively. For third-party recipient classification (P6), we achieve a 0.95 F1-score for the named and categorized third-party categories, and 0.89 for the public/other user category.

Prior work [34, 64] evaluated on OPP-115 using its fine-grained attribute labels, training separate classifiers for the main categories (e.g., first-party collection vs. third-party sharing) and the fine-grained attributes (e.g., PI categories). While we evaluate using the same dataset, our task is distinct because we categorize PI categories and their context (first-party collection vs third-party sharing) together as one task. Using two separate classifiers produces ambiguities/inaccuracies in real-world measurements, such as if a policy segment discusses both contexts and PI categories get mixed across the two contexts.

Here we discuss prior work to illustrate the relative performance of our approach, given the lack of equivalent baselines. Polisis [34] reported an average F1-score of 0.81 for PI categories using a CNN, and Wagner [64] reported an average F1-score of 0.828 using BERT-based models, both under a mixed-context settings. In comparison, we achieve an average F1-score of 0.88 for first-party collection (P1) and 0.89 for third-party collection (P2). We note that there are certain personal information types for which our F1-scores are lower than those reported in prior work. For example, for contact information, Polisis reported an F1-score of 0.90, Wagner reported 0.965, and we achieve 0.92 for first-party collection and 0.81 for third-party sharing. Again, we note that our task is different from prior work, as we classify the attributes along with their context.

Next, since we introduce four new PI classes, and refine the definition of "identifier" classes (Section 4.2), we further reproduce a prior approach to compare its performance with our's on these classes. While Polisis and Wagner's work did not release source code for their approaches, prior work [9] has attempted to reproduce Polisis's results. We therefore adopt the implementation from this open-source repository for comparison (i.e., using a CNN), using 70% of the labeled segments for training and 30% for testing, with balanced positive and negative samples and a mixed-context setting. We find that our approach outperforms Polisis across these personal information types. For identifiers, we achieve an F1-score of 0.88 for first-party collection and 0.93 for third-party sharing, compared to 0.82 using Polisis. For inferred data, our scores are 0.91 (first-party) and 0.89 (third-party), versus 0.85 for Polisis. For sensor data, our scores are 0.97 (first-party) and 1.00 (third-party), compared to 0.94 for Polisis. For other sensitive information, we achieve scores of 0.96 (first-party) and 0.97 (third-party), while using Polisis achieves 0.88. On biometric data, both our method and Polisis achieve 1.00.

For data purpose categories, we achieve average F1-scores of 0.86 for first-party (P4) and 0.84 for third-party (P5), exceeding or matching the mixed-context evaluation reported by Polisis (0.83) and Wagner's (0.84). Polisis did not cover third-party recipients (P6), and Wagner reported F1-scores of 0.825 (named/categorized third-party categories) and 0.750 (public), compared to our scores of 0.95 and 0.89, respectively.

Overall, although the inspection tasks are more complex and detailed than the coverage tasks, our LLM method achieved reasonable performance. Even when compared to prior non-equivalent tasks, our approach overall outperforms existing methods, and our classifications are also broader (see Section 4.2). Thus, this approach enables us to apply the analysis framework for meaningful large-scale measurements.

5 Privacy Policy Collection

In this section, we describe the process of collecting our privacy policy dataset for measurement, including how we se-

Personal Information	1st (P1)	3rd (P2)
Financial/Commercial	0.84	0.88
Health	0.98	0.97
Biometric	1.00*	1.00*
Contact	0.92	0.81
Geolocation	0.84	0.89
Demographic	0.90	0.81
Identifier	0.88*	0.93*
User online activities	0.82	0.85
User profile	0.73	0.85*
Social media data	0.81	0.92*
IP address and device IDs	0.89	0.83
Cookies and tracking elements	0.94	0.97
Computer information	0.80	0.82
Survey data	0.84	0.81
Inferred Data	0.91*	0.89*
Sensor	0.97*	1.00*
Other sensitive information	0.96*	0.97*
Average	0.88	0.89

Table 4: F1-scores from evaluating our method's classification of personal information category (for both first-party collection and third-party disclosure). Values with * are evaluated on our dataset.

lected target domains, crawled web pages, and identified their privacy policies.

Target Domains. We selected popular domains from the Chrome User Experience Report (CrUX) [30] top list, as it is widely used and considered a more accurate representation of website usage compared to other DNS-based top lists, such as SecRank [70], the Umbrella top list [19], or the aggregated Tranco list [54]. As discussed in Section 3, US state privacy laws typically define their scope of applicability based on business size. Therefore, we focus on the top 100K domains from the CrUX list, as these are more likely to fall within the scope of such laws. For similar considerations, we also include the domains of the top 1K US companies [27] and the top 500 EU companies [28], as listed by Fortune 2024. This also enables us to study the differences in privacy policy content across companies from the EU and the US.

Data Collection Process. We collected privacy policies for the target domains in August 2024, using the following steps:

1) **Finding candidate privacy policies**. Privacy policy links are not universally standardized in their text or hyperlink format. We identify hyperlinks presumed to point to privacy policies for further evaluation on each domain's homepage through keyword matching (akin to the methods used by prior work on web privacy policy collection [8, 22, 46]). We use GDPR-related keywords from existing multilingual word lists from [22] and those specific to CCPA.
2) **Crawling web pages**. For each homepage and its candidate privacy policy pages, we use Selenium to load each page, waiting up to 30 seconds for it to load and then downloading the HTML file. The traffic load generated by our crawling on each website should be negligible.
3) **Text Extraction.** Next, we removed boilerplate content (e.g., script, style, head tags, etc.) and extracted the main content from the HTML page into Markdown format, using the html-to-markdown converter in Go [39], the same library used by the Firecrawl service [2]. Although the readability.js library is commonly used for text extraction in prior work [8, 35], our tests revealed that both readability.js as well as Chrome's new reading mode perform poorly with HTML tables, often returning blank outputs. We observed that many privacy policies now use tables for information disclosure as required by CCPA. Thus, using the Markdown format as LLM inputs helps maintain the table context and the overall structure of the privacy policy.
4) **Language detection**. Given that our study is motivated by GDPR and the US state privacy laws, and that analyzing multilingual content in the EU requires significant effort and additional language expertise, we limit our focus to English-language privacy policies. To ensure high accuracy in language detection, we adopt the method from [35], which runs eight open-source libraries for language detection on each web page and performs majority voting on their outputs. We then excluded domains whose homepage is non-English or where all its candidate privacy policy pages are non-English at this step.
5) **Privacy policy classification.** Finally, for domains with at least one candidate privacy policy page, we use a pre-trained machine learning classifier from prior work [35], which reports a 99.1% accuracy in detecting English privacy policies, to classify each candidate page.
6) **Selecting privacy policies with high-confidence.** Since our measurement objective is strongly correlated with crawling the correct privacy policy, we apply two considerations for further selecting high-confidence privacy policies. First, we observed that some terms of service or disclaimer pages may contain privacy-related content, leading to misclassification. Therefore, we further select pages identified by

Purposes	1st (P4)	3rd (P5)
Performing services	0.89	0.85
Advertising/Marketing	0.82	0.90
Analytics/Research	0.90	0.87
Personalization/Customization	0.78	0.54
Service Operation and Security	0.82	0.89*
Legal requirement	0.83	0.85
Merger/Acquisition	0.95*	0.97
Average	0.86	0.84

Table 5: F1-scores from evaluating our method's classification of first-party collection and third-party disclosure purposes. Values with * are evaluated on our dataset.

Cases	Total	1K	10K	100K	EU	US
High Confidence	18,442	246	1,709	15,677	231	760
Plausible	6,684	95	689	5,681	105	141
No Policy Link	22,311	180	1,826	20,153	79	98
Non-English	48,178	445	4,364	43,307	79	1
Access Failure	2,578	16	174	2,382	6	0
Blank Homepage	3,056	18	238	2800	0	0
Full Set	101,249	1K	9K	90K	500	1K

Table 6: Distribution of the number of domains in different detection cases across the CrUX top 1K, 10K (excl. 1K), 100K (excl. 10K), and Fortune EU top 500 and US top 1K.

the classifier as privacy policies, where the URL contains keywords such as "privacy", and manually check all privacy policies crawled using CCPA-specific keywords.
Second, as the classifier only determines whether a page is a privacy policy or not, we need to confirm that the privacy policy belongs to the targeted domain to ensure measurement accuracy. We notice a common corner case that a candidate privacy policy's second-level domain (SLD) may differ from the targeted domain's, which can be either problematic or not. For instance, a domain using a Google service might include Google's privacy policy link on its page (false positive), while a domain's privacy policy page might link to its parent company's policy with a different SLD (true positive). To automate the selection process, which affords large-scale evaluations than relying on manual analysis, we streamline the process by selecting 1) candidates whose SLD matches the target domain, and 2) candidates whose SLD matches any domain the target redirects to, as recorded during crawling.

Table 6 lists the number of domains with high-confidence privacy policy pages found for each domain group, along with the distribution of different causes for excluding domains at each processing step. In the table, "No Policy Link" refers to cases where no candidate link is found on the page, or all candidate pages are classified as non-privacy policy. "Blank Homepage" also includes cases where the homepage follows a specific pattern without any candidate links found, such as the default Nginx welcome page.

We manually checked 100 randomly selected domains in the plausible group and obtained 76% accuracy, with most inaccuracies arising from privacy policies that do not belong to the target domain, as discussed. We then manually checked 100 randomly selected domains in the high-confidence group, reaching 99% accuracy, with one error being a privacy center (navigation) page that links to different privacy policies. To prevent such cases, we manually reviewed and corrected all privacy policies with few words in the dataset.

We use domains and their privacy policies from the high-confidence set as our test dataset for measurement in Section 7. From the top 100K group, we randomly selected 2K domains for measurement to ensure a representative yet computationally efficient subset for analysis. For other domain groups, we use the full high-confidence set. In total, our test dataset consists of 4,896 distinct domains and 5,210 unique privacy policy documents (some domains have multiple policies, such as a main policy and a CCPA-specific one).

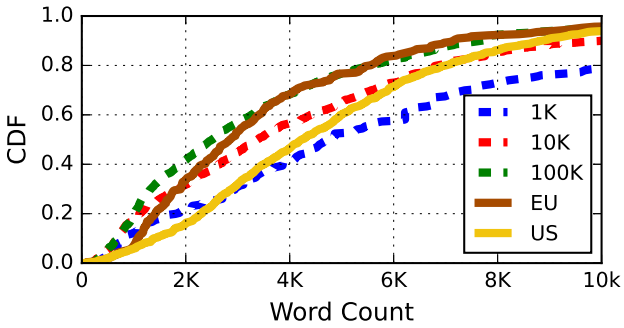

Figure 1: CDF of the word count of privacy policies across different domain groups, with the x-axis truncated at 10K.

6 Limitations

Before presenting our large-scale measurement results, we discuss the limitations of our measurement and analysis.

Our work focuses on systematizing Western privacy laws (Section 3) and analyzing English privacy policies (Section 5), which may not generalize to other regions. Also, as discussed in Section 3, many clauses may not apply to a domain's privacy policy, depending on the various characteristics of the online service and its users. We do not evaluate policy compliance with privacy laws and leave determining which privacy laws apply to which domains for future work. Instead of investigating compliance, we focus on comprehensively characterizing what privacy practices are currently disclosed via privacy policies. Finally, our LLM-based analysis framework is not perfectly accurate (likewise with all prior automated approaches), so our results have some margin of error. However, our method exhibits overall high accuracy, so we believe our findings still provide meaningful insights into the privacy practices discussed in real-world privacy policies.

7 Large-Scale Privacy Policy Measurement

We now apply our analysis framework to the privacy policies of 4,896 domains (see Section 5).

7.1 Policy Length and Readability

We begin with simple characterizations of our privacy policies, specifically their length and readability. Figure 1 shows the CDF of policy length, as measured by word count, across

each domain group. We see that higher-ranked domains tend to provide longer privacy policies. Regionally, US domains feature longer privacy policies than those in the EU, possibly in response to more detailed requirements in US state privacy laws (see Section 3). We also observe substantial variance in word counts, as reflected in the CDF. The longest privacy policy in our dataset is Microsoft's, totaling 50,409 words.

Prior work [8] found that the Alexa top domains' privacy policies doubled in median word count from 2009 to 2019, with a sharper rise after GDPR. Compared with their findings, we observe a continued increase in privacy policy word counts. The median word count in our dataset is 4,820 for the top 1K domains, 3,389 for the top 10K, and 2,512 for the top 100K, compared to 3.1K, 2.4K, and 1.7K in Alexa's top 1K, 10K, and 100K domains in 2019 [8], respectively.

Several prior studies [24, 48] have found that privacy policies are difficult to read. Similar to [8], we measure the readability using the Flesch-Kincaid Grade Level (FKGL). The median FKGL of privacy policies in our dataset is 15.11, representing a college level of education necessary for comprehension. We note this FKGL score is higher than past reports from 2012 (13.33 [44]), and in 2019 (13.2 [8]). Thus, privacy policies are becoming longer and more challenging for consumers to understand.

7.2 Coverage of Privacy Policy Content

Here, we evaluate the extent to which our privacy policies cover the privacy practice clauses discussed across the ten privacy laws (through our coverage tasks). For each clause, Table 7 lists the percent of domains that include the required content in their privacy policies, organized by domain groups.

Personal Information Practices. We see that privacy policies heavily focus on personal information practices, with over 90% of all domains discussing P1-5 (PI collection + disclosure). P7 (PI retention) exhibits lower inclusion (74% overall), likely because only GDPR requires this clause.

Consumer Privacy Rights. We observe that privacy policies discuss consumer rights less often than personal information practices and that there's wide variation across rights-related clauses. The right to know (R1), delete (R3), correct (R4), and opt-out of ads/sales/profiling (R5) are disclosed by over 75% of all policies. Other rights are discussed by 57% or fewer of the policies. The right to non-discrimination (R10) is specific to California's CCPA and the least included overall.

For methods of exercising rights, we see majority coverage only for submitting privacy requests (E1, 82% of policies) and contact information (E3, 90% of policies). For the other clauses on exercising rights (notably including those about opt-out signals), only a minority of domains' privacy policies provide information. (E6, on opting into data disclosure for those under 16 years of age, is particularly low, but only applies to businesses that collect data from this population.)

Cl.	Total	1K	10K	100K	EU	US
P1	97%	98%	96%	97%	94%	99%
P2	91%	98%	92%	89%	85%	94%
P3	94%	97%	93%	94%	91%	96%
P4	98%	99%	97%	98%	98%	99%
P5	95%	99%	95%	94%	92%	96%
P6	94%	96%	93%	93%	96%	95%
P7	74%	85%	75%	68%	90%	82%
R1	79%	87%	79%	75%	94%	88%
R2	55%	70%	56%	48%	83%	61%
R3	80%	87%	80%	76%	93%	89%
R4	78%	80%	78%	74%	93%	87%
R5	76%	82%	77%	71%	66%	88%
R6	48%	46%	47%	45%	84%	52%
R7	57%	65%	57%	49%	76%	70%
R8	57%	69%	59%	49%	87%	60%
R9	56%	69%	59%	49%	78%	56%
R10	34%	44%	34%	26%	11%	61%
E1	82%	88%	82%	78%	90%	91%
E2	22%	33%	23%	16%	10%	37%
E3	90%	93%	89%	86%	96%	97%
E4	49%	61%	51%	39%	41%	70%
E5	30%	40%	29%	22%	10%	57%
E6	3%	4%	3%	3%	2%	3%
E7	31%	30%	31%	26%	13%	54%
E8	13%	13%	12%	10%	5%	30%
D1	7	0	4	0	0	4
D2	1	0	1	0	0	0
D3	22%	33%	23%	15%	9%	44%
D4	19%	17%	17%	13%	9%	45%
D5	70%	79%	69%	63%	70%	88%
D6	7%	27%	10%	4%	1%	8%
D7	94%	98%	94%	93%	95%	95%
D8	35%	49%	38%	28%	49%	43%
D9	30%	37%	31%	25%	44%	36%

Table 7: Percentage of clause coverage across different domain groups.

Specific Disclosures. We detect limited coverage of specific disclosure clauses, except for the last updated date (D5, 70%) and D7 (third-country PI transfer, 94%). D7 likely exhibits widespread adoption as it is similar to personal information practices, which privacy policies discuss more extensively (although we classify as a Disclosure clause given it entails a specific scenario).

Ranking Differences. We find a clear trend that higher-ranked domains cover more clauses in their privacy policies, for all clauses. We expect that higher-ranked domains are more likely to fall under applicable privacy laws, requiring them to provide more information in their privacy policies. In many cases, they also are more likely to handle user data (and thus have more privacy practices to disclose). D6 (metrics reports on consumer privacy requests) illustrates this as it only applies to businesses collecting large amounts of user data, and exhibits a 17% coverage gap between the top 1K and top 10K, and a 23% gap with the top 100K.

Regional Differences. Clause coverage can differ notably

No.	Personal Information	P1	P2
1	Identifier	96%	90%
2	Contact	92%	83%
3	Cookies and Tracking Elements	89%	78%
4	IP Address and Device IDs	87%	72%
5	User Online Activities	87%	78%
6	Computer Information	81%	58%
7	User Profile	70%	45%
8	Demographic	63%	45%
9	Geolocation	63%	47%
10	Financial/Commercial	61%	52%
11	Inferred Data	43%	24%
12	Survey Data	32%	15%
13	Social Media Data	31%	35%
14	Other Sensitive Information	27%	17%
15	Health	20%	13%
16	Sensor	17%	13%
17	Biometric	16%	10%

Table 8: Among domains collecting or sharing personal information, we list the percent that collect (P1) or share (P2) each personal information type (sorted by P1).

between top US and EU domains, demonstrating the influence of regional regulations despite the inter-connected nature of the web. For example, R10 (right to non-discrimination) is CCPA-specific and is in the policies of only 11% of top EU domains compared to 61% of top US domains. Overall, top US domains align closer with CCPA-specific requirements (and other US states to a lesser extent), whereas top EU domains align closer to GDPR-exclusive clauses.

7.3 Personal Information Practices

Here, we analyze the personal information practices identified by our inspection tasks, specifically the PI types collected and shared (P1, P2), collection and sharing purposes (P4, P5), and third-party recipients of PI (P6).

Categories of Personal Information. Among domains disclosing the collection (P1) or sharing (P2) of PI, Table 8 lists the percent of domains collecting or sharing (including selling/disclosing) each PI category. We observe broad and diverse collection and sharing of consumer PI, particularly consumer contacts, cookies/trackers, IP addresses, user online activity, and computer information (e.g., OS version). For nearly all PI classes, the majority of domains that collect such PI also share it, highlighting the privacy consequences of even first-party data collection. Interestingly, the top category described for both collection and sharing is "identifier", which is an expansive and generically defined category (see Section 4.2). Thus, users are often unlikely to understand concretely the personal identifiers collected and shared. We also note that a majority of domains collect geolocation (of which 75% also share), emphasizing the importance of location privacy. Furthermore, our analysis includes new categories of PI compared to prior work [67], including biometric, sensor, other sensitive data, and inferred data, which are collected by 16–43% of the domains (and shared by 10–24%).

No.	Business/Commercial Purposes	P4	P5
1	Performing Services	97%	89%
2	Service Operation and Security	96%	77%
3	Analytics/Research	89%	76%
4	Advertising/Marketing	85%	82%
5	Legal Requirement	83%	82%
6	Personalization/Customization	77%	47%
7	Merger/Acquisition	27%	56%

Table 9: Among domains collecting or sharing personal information, we list the percent that discuss a purpose for collecting (P4) or sharing (P5), for each class of purposes (sorted by P4).

Business/Commercial Purposes. Table 9 lists the percent of domain privacy policies that discuss different business/commercial purposes for collecting (P4) and sharing/selling/disclosing (P5) consumer PI. For both collection and sharing, nearly all purposes are described by the vast majority (over 75%) of domain privacy policies, except for merger/acquisition purposes, and personalization/customization specifically for PI sharing. This observation signals that consumer PI is broadly used for multiple diverse purposes. We also see that over 95% of domains collecting for advertising/marketing purposes are also sharing for the same purpose, demonstrating the extensive influence of third-party advertising. Similarly, over 80% of domains collecting for security are also sharing for this reason, suggesting widespread use of third-party security solutions. Finally, nearly all domains that collect for legal requirements also share likewise, indicating that the legal requirements themselves entail data sharing with government or legal authorities.

Third-Party PI Recipients. Here, we analyze popular third-party PI recipients (P6). As discussed in Section 4.2, our LLM method distinguishes between named and categorized entities. We manually merged identified named entities (e.g., "Google" and "Google LLC.") and categories (such as "payment processors" and "payment service providers"), and list the top 10 for each in Table 10. Note that these percentages represent lower bounds on how many domains engage with a particular entity, as a named entity in one domain's policy may be a categorized entity in another (and vice versa).

Among named entities, Google services are particularly dominant, accounting for 5 of the top 10 third-party PI recipients. In total, 38% of domains discuss one of the Google services as a PI recipient. Other social media platforms, such as Facebook, Twitter, and LinkedIn, are also commonly named entities, as well as advertising networks (NAI, DAA). The distribution of third-party categories aligns with that of named entities, reinforcing the observation of extensive third-party

No.	Named	%	Categorized	%
1	Google	26%	Advertising	49%
2	Facebook	17%	Affiliates	33%
3	G-Analytics	16%	Government	29%
4	Twitter	8%	Analytics/Measurement	23%
5	Youtube	4%	LEAs[3]	19%
6	Google Ads	4%	Marketing	16%
7	LinkedIn	3%	Social Networks	16%
8	NAI[1]	3%	Payment Processors	14%
9	DAA[2]	3%	Vendor	13%
10	DoubleClick	2%	Contractors	9%

[1] NAI: Network Advertising Initiative
[2] DAA: Digital Advertising Alliance
[3] LEAs: Law Enforcement / Authorities

Table 10: The top 10 named and categorized third-party PI recipients, ranked by the percent of domain privacy policies disclosing the recipient entities.

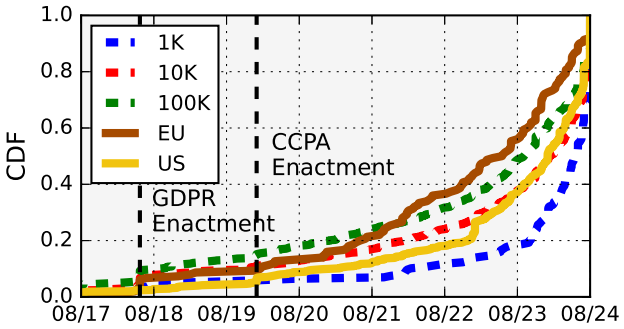

Figure 2: CDF of privacy policy last update dates across different domain groups, with the x-axis truncated from 08/2017 to 08/2024 (the time of our privacy policy collection).

advertising/marketing. Similar to PI collection and sharing purposes, we find a notable portion of domains sharing data with government (29%) and law enforcement (19%).

7.4 Case Studies

From our coverage tasks, our LLM extracted policy text relevant to each clause, affording deeper analysis. Here, we conduct several case studies of our collected data, providing broader insights into privacy policies and illustrating the value of an LLM-based privacy policy analysis framework.

7.4.1 Last Updated Dates

We identified 3,414 domains (70%) that provide a last updated date in their privacy policies, aligning with the CCPA requirement (D5). Here, we further prompt our LLM to extract and parse the last updated date in a standard format. Figure 2 plots the distribution of these dates for different domain groups.

We find that most domains (>80%) have updated their privacy policies since the enactment of GDPR and CCPA, reflecting an overall trend of recent updates to privacy policies. More popular domains tend to have more recently updated policies, and likewise, the top US companies are more recently revised compared to the top EU ones. We do see that a non-trivial minority of domains have not updated their policy within one year of our data collection, though, with the most extreme example being dated to 2002 (Wabash National[8]). Interestingly, we also found sites with invalid update dates, such as "31.02.2022[9]". While some privacy laws (e.g., Florida's) require frequent updates (e.g., annually), we note that such frequency may challenge consumers when tracking changes and monitoring privacy practices.

7.4.2 Types of Contacts

While most privacy laws require an organization's contact (E3), Connecticut and Oregon's privacy laws specifically require an online contact (e.g., email address, online form), presumably for ease of access. We observe contact information provided by over 90% (4,390) of our domains' privacy policies. We then further prompt our LLM to classify the contact details extracted as online versus offline contact methods.

Among domains with contacts, 98% offer online methods, with 82% supporting an email address. We also find 37% providing a phone number, although this is not a strictly online contact. Overall, 63% offer both online and offline contacts (e.g., physical address), while only 90 domains (2%) provide offline options exclusively[10]. Thus, most domains already support online contacts (possibly guided by privacy laws).

7.4.3 Frictionless Opt-Out Signals

California's CCPA requires that if a business handles frictionless opt-out preference signals, the privacy policy explains how consumers can implement those signals (E8). We observed only 13% (639) domains with this clause, so such frictionless opt-out signals are not widely discussed.

Using the policy text for the E8 clause extracted by our LLM, we next use keyword matching to explore how these domains currently support frictionless opt-out. We found that the vast majority of these domains (506, or 79%) mentioned Global Privacy Control, or GPC [3], as a frictionless opt-out signal. GPC is a recent effort to create a standardized opt-out signal for online privacy, and is available on some browsers (e.g., Firefox, Brave), or via a browser extension. Regionally, we observe significantly more top US domains (191) respecting GPC, compared to only 10 top EU domains,

[8] https://web.archive.org/web/20250114183052/https://www.onewabash.com/about-us/privacy-policy
[9] https://web.archive.org/web/20241229102121/https://www.sinsay.com/sq/en/privacy-policy
[10] https://web.archive.org/web/20250110115302/https://www.eqt.com/privacy-policy

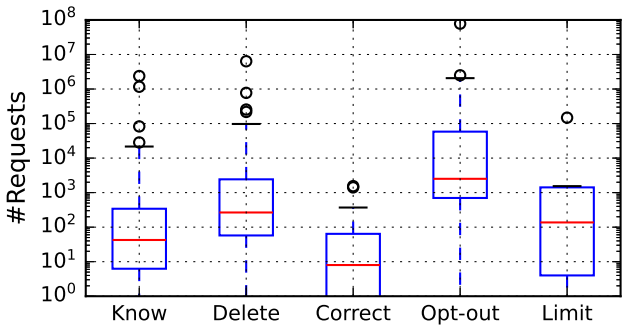

Figure 3: Box plots of the number of consumer privacy requests received by a domain for each right in the past calendar year (whiskers represent the 5th and 95th percentiles).

reflecting the CCPA-specific nature of this clause. Beyond GPC, we observe alternative signals such as Do Not Track (DNT) or HTTP headers, which are not mandatory.

7.4.4 Metrics Report on Consumers' Privacy Requests

We observe that the privacy policies of 359 domains provided the metrics report about consumer privacy requests (D6), which details the number of different types of privacy requests received during the past calendar year and the average response time for those requests. For 68 domains, the report was directly embedded in the policy document (whereas others were linked to an external page). We manually analyzed these reports and plotted the distribution of report metrics in Figure 3.

We find that domains generally receive a limited number of requests per year, with the median ranging from 42 (right to know) to 266 (right to delete) requests. Opt-out requests are the exception, with a significantly higher median of 2,511 requests. We do observe domains receiving large numbers of requests in a year, though; Spotify and T-Mobile received 78M and 22M opt-out requests, respectively (though, we note these companies also have massive consumer bases).

For request response times, we find that domains largely prioritize opt-out requests, which have a median response time of 1 day, compared to 3.5 days for correction requests and over 10 days for other requests. Furthermore, CCPA mandates a maximum response time of 90 days for consumer requests, and we observed no violations in those metric reports (although only the average response time is reported).

8 Concluding Discussion

In this work, we developed an LLM-based automated approach for comprehensively evaluating web privacy policies, grounded in our systematization of privacy clauses in 10 modern privacy laws. We then applied our LLM method to investigate the privacy practices disclosed in real-world privacy policies at scale. Here, we synthesize lessons, recommendations, and future directions from our study for different stakeholders in the web privacy space.

Websites/Businesses. Overall, we observed that privacy policies are becoming longer and increasingly difficult for consumers to understand (Section 7.1) (echoing prior work [24, 48]). Modern privacy policies also include diverse content under various laws (Section 7.2), particularly in higher-ranking and US domains. Meanwhile, we find that privacy policies continue shifting with recent updates (Section 7.4.1). Regularly tracking these updates to privacy policies further challenges consumers. Thus, we recommend that websites explore ways to simplify their privacy policies, surface privacy practices to users in a digestible fashion (while maintaining a full privacy policy document), and notify consumers with clear, concise lists of changes when policy updates occur.

In Section 7.2, we also found that privacy policies are less likely to provide information on consumer privacy rights and methods for exercising them, compared to personal information practices. We recommend that the website provide broader visibility into and support of consumer privacy rights. Related, we found limited discussed support of frictionless opt-out signals in Section 7.4.3. Domains could more broadly implement processing of such signals, including GPC [3].

Consumers. We observed that privacy policies disclose extensive user data collection and sharing (Section 7.2) and that data sharing exhibits concentration among common third-party entities (Section 7.3). Thus, consumers can actively engage with the policies of both websites they use, as well as the third-party entities receiving personal information, to better understand and control their data, especially as many do describe consumer privacy rights and methods for exercising those rights. Overall though, we have observed limited use of these rights, given the low numbers of privacy requests that organizations have reported receiving (Section 7.4.4). Thus, user education could prove impactful in raising awareness about actionable privacy rights.

Regulators. We validate in this work that privacy laws have indeed impacted privacy policies and privacy practices. In certain cases (although more limited generally), we have seen millions of users exercising their privacy rights to opt out of the sale and sharing of their personal information (Section 7.4.4). We have also seen that the majority of policies have been updated since landmark regulations (Section 7.4.1). Furthermore, we observed regional differences in privacy practices disclosed, largely aligning with the local regulations (Section 7.2). Thus, we are optimistic that future developments in the privacy regulation space can strengthen consumer privacy protections.

We did observe some challenges with existing privacy laws, though, where certain definitions are overly broad or generic. For example, in Section 7.2, we found that the "identifier" per-

sonal information category is the most commonly discussed in privacy policies as collected and/or shared. Yet, its definition (in CCPA) is expansive, leaving users (and possibly websites) uncertain about what types of data are concretely considered. Therefore, regulators should prioritize narrowly-scoped category definitions for greater clarity.

Researchers and Future Work. Our findings demonstrate the need to consider different privacy laws when analyzing privacy policies, as laws vary in the types of privacy practices considered (Section 3). In Section 7.2, we saw that clauses specific to certain laws (particularly GDPR and CCPA) do exhibit high adoption, particularly regionally. Even for clauses exclusive to relatively new laws (e.g., D1 and D2, in Texas and Florida's privacy laws), we observe some domains already including these clauses in their policies. Therefore, comprehensively understanding a website or business's privacy practices requires accounting for multiple regulations.

Meanwhile, in our inspection tasks, we needed to refine the definitions of personal information practices beyond those of prior work [67]. Section 7.3 found that 16%–43% of domains collect the new types of personal information that we incorporated. Future research can explore new personal information produced by emerging technologies, such as IoT devices (and other cyber-physical systems) and augmented/virtual reality.

We showed that LLMs enable more comprehensive, accurate, and flexible analysis of privacy policies, not only for assessing clause coverage but also by examining the LLM text extracted for clauses (as seen in the case studies of Section 7.4). LLMs have text analysis capabilities beyond English, and future work can explore non-Western privacy laws as well as develop pipelines for multilingual privacy policy analysis. There are human-annotated privacy policy corpora in non-English languages, such as German [13] and Chinese [66, 73], and future research may evaluate LLM performance using these existing corpora or create new ones for more languages. Finally, future work can also build upon our dataset and analysis pipeline to more deeply explore web privacy practices, such as analyzing contradictions/inconsistencies in practices, and comparing practices with websites' actual data collection behaviors.

9 Ethics Considerations

Our study's primary ethical consideration pertains to our web crawling of website privacy policies. To ensure that our crawling has minimal impact on websites, we rate-limited our crawl of sites, such that we do not crawl a site faster than one page every 30 seconds (in most cases, much slower). Such a rate should not burden even small websites (and our study population consists of top websites). Note that the data we collect from sites is public information (i.e., privacy policies). As part of the study, we do not assert that any of the privacy policies violate the proposed clauses under different privacy laws. Hence, the end users are not impacted by our study. We do not believe there are ethical considerations with our subsequent analysis of this data.

10 Open Science

We provide open access to our full study's dataset collection and analysis methods at https://github.com/BEESLab/LLM-PP2025[11], including the privacy policy web crawler, the collected dataset of privacy policies, the LLM-based assessment framework, and the produced LLM-based policy annotations for both the coverage and inspection tasks (the LLM-PP2025 dataset).

Acknowledgments

This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA. The authors also thank the anonymous reviewers for their valuable feedback.

References

[1] California Code, CIV 1798.140. https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=1798.140., 2024.

[2] Firecrawl. https://www.firecrawl.dev/, 2024.

[3] Global Privacy Control. https://w3c.github.io/gpc, 2024.

[4] Introducing Llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/, 2024.

[5] Llama-3.1-405B. https://huggingface.co/meta-llama/Llama-3.1-405B, 2024.

[6] Llama-3.1-70B. https://huggingface.co/meta-llama/Llama-3.1-70B, 2024.

[7] Andrick Adhikari, Sanchari Das, and Rinku Dewri. PolicyPulse: Precision semantic role extraction for enhanced privacy policy comprehension. In *the Network and Distributed System Security Symposium (NDSS)*, 2025.

[8] Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. Privacy policies over time: Curation and analysis of a million-document dataset. In *the Web Conference (WWW)*, 2021.

[11] Our artifacts are also publicly available via the permanent Zenodo link: https://zenodo.org/records/15594020.

[9] Smart Data Analytics. Polisis Benchmark. https://github.com/SmartDataAnalytics/Polisis_Benchmark, 2020.

[10] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. PolicyLint: Investigating Internal privacy policy contradictions on Google Play. In *USENIX Security Symposium (USENIX Security)*, 2019.

[11] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. Actions speak louder than words: Entity-Sensitive privacy policy and data flow analysis with PoliCheck. In *USENIX Security Symposium (USENIX Security)*, 2020.

[12] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[13] Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar K Bannihatti, Tristan Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, et al. A tale of two regulatory regimes: Creation and analysis of a bilingual privacy policy corpus. In *the Language Resources and Evaluation Conference (LREC)*, 2022.

[14] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, et al. Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text. In *the Web Conference (WWW)*, 2020.

[15] Veronika Belcheva, Tatiana Ermakova, and Benjamin Fabian. Understanding website privacy policies—a longitudinal analysis using natural language processing. *Information*, 14(11):622, 2023.

[16] Duc Bui, Brian Tang, and Kang G Shin. Do opt-outs really opt me out? In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022.

[17] Duc Bui, Yuan Yao, Kang G. Shin, Jong-Min Choi, and Junbum Shin. Consistency analysis of data-usage purposes in mobile apps. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2021.

[18] Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-jun Li, and Yaxing Yao. Clear: Towards contextual llm-empowered privacy policy analysis and risk generation for large language model applications. In *Annual ACM Conference on Intelligent User Interfaces (IUI)*, 2025.

[19] Cisco. Umbrella Popularity List . https://s3-us-west-1.amazonaws.com/umbrella-static/index.html, 2024.

[20] Hao Cui, Rahmadi Trimananda, Athina Markopoulou, and Scott Jordan. PoliGraph: Automated privacy policy analysis using knowledge graphs. In *USENIX Security Symposium (USENIX Security)*, 2023.

[21] DataGuidance. Comply with US Privacy Laws. https://www.dataguidance.com/topics/privacy-overview?topic=usprivacylaws, 2024.

[22] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy. In *the Network and Distributed System Security Symposium (NDSS)*, 2019.

[23] Jose M Del Alamo, Danny S Guaman, Boni García, and Ana Diez. A systematic mapping study on automated analysis of privacy policies. *Computing*, 104(9), 2022.

[24] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. Large-scale readability analysis of privacy policies. In *the International Conference on Web Intelligence (WIC)*, 2017.

[25] Mafalda Ferreira, Tiago Brito, José Fragoso Santos, and Nuno Santos. RuleKeeper: GDPR-aware personal data compliance for web frameworks. In *IEEE Symposium on Security and Privacy (S&P)*, 2023.

[26] Olivia Figueira, Rahmadi Trimananda, Athina Markopoulou, and Scott Jordan. Diffaudit: Auditing privacy practices of online services for children and adolescents. In *the ACM Internet Measurement Conference (IMC)*, 2024.

[27] Fortune. Fortune 500. https://fortune.com/ranking/fortune500/search, 2024.

[28] Fortune. Fortune 500 Europe. https://fortune.com/europe/ranking/fortune500-europe/, 2024.

[29] Arda Goknil, Femke B Gelderblom, Simeon Tverdal, Shukun Tokas, and Hui Song. Privacy policy analysis through prompt engineering for LLMs. *arXiv preprint arXiv:2409.14879*, 2024.

[30] Google. Adding Rank Magnitude to the CrUX Report in BigQuery. https://developer.chrome.com/blog/crux-rank-magnitude, 2021.

[31] Elias Grünewald and Frank Pallas. Tilt: A gdpr-aligned transparency information language and toolkit for practical privacy engineering. In *the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[32] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems (NIPS)*, 2024.

[33] Rajaa El Hamdani, Majd Mustapha, David Restrepo Amariles, Aurore Troussel, Sébastien Meeùs, and Katsiaryna Krasnashchok. A combined rule-based and machine learning approach for automated GDPR compliance checking. In *the International Conference on Artificial Intelligence and Law (ICAIL)*, 2021.

[34] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *USENIX Security Symposium (USENIX Security)*, 2018.

[35] Henry Hosseini, Martin Degeling, Christine Utz, and Thomas Hupperich. Unifying privacy policy detection. In *Privacy Enhancing Technologies Symposium (PETS)*, 2021.

[36] Henry Hosseini, Christine Utz, Martin Degeling, and Thomas Hupperich. A bilingual longitudinal analysis of privacy policies measuring the impacts of the GDPR and the CCPA/CPRA. In *Privacy Enhancing Technologies Symposium (PETS)*, 2024.

[37] Umar Iqbal, Pouneh Nikkhah Bahrami, Rahmadi Trimananda, Hao Cui, Alexander Gamero-Garrido, Daniel J Dubois, David Choffnes, Athina Markopoulou, Franziska Roesner, and Zubair Shafiq. Tracking, profiling, and ad targeting in the alexa echo smart speaker ecosystem. In *the ACM Internet Measurement Conference (IMC)*, 2023.

[38] Kaushal Kafle, Prianka Mandal, Kapil Singh, Benjamin Andow, and Adwait Nadkarni. Understanding the privacy practices of political campaigns: A perspective from the 2020 us election websites. In *IEEE Symposium on Security and Privacy (S&P)*, 2024.

[39] Johannes Kaufmann. html-to-markdown. https://github.com/JohannesKaufmann/html-to-markdown, 2024.

[40] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NIPS)*, 2022.

[41] Noam Kolt. Predicting consumer contracts. *Berkeley Tech. LJ*, 2022.

[42] Lin Kyi, Sushil Ammanaghatta Shivakumar, Cristiana Teixeira Santos, Franziska Roesner, Frederike Zufall, and Asia J Biega. Investigating deceptive design in gdpr's legitimate interest. In *CHI Conference on Human Factors in Computing Systems (CHI)*, 2023.

[43] Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[44] Yuanxiang Li. Online privacy policy of the thirty Dow Jones corporations: Compliance with FTC Fair information practice principles and readability assessment. *Theses Digitization Project*, 2012.

[45] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. The privacy policy landscape after the GDPR. In *Privacy Enhancing Technologies Symposium (PETS)*, 2020.

[46] Shuang Liu, Baiyang Zhao, Renjie Guo, Guozhu Meng, Fan Zhang, and Meishan Zhang. Have you been properly notified? Automatic compliance analysis of privacy policy text with GDPR article 13. In *the Web Conference (WWW)*, 2021.

[47] Sunil Manandhar, Kapil Singh, and Adwait Nadkarni. Towards automated regulation analysis for effective privacy compliance. In *the Network and Distributed System Security Symposium (NDSS)*, 2024.

[48] George R Milne, Mary J Culnan, and Henry Greene. A longitudinal assessment of online privacy notice readability. *Journal of Public Policy & Marketing*, 25(2):238–249, 2006.

[49] Maaz Bin Musa, Steven M Winston, Garrison Allen, Jacob Schiller, Kevin Moore, Sean Quick, Johnathan Melvin, Padmini Srinivasan, Mihailis E Diamantis, and Rishab Nithyanand. C3PA: An open dataset of expert-annotated and regulation-aware privacy policies to enable scalable regulatory compliance audits. In *the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[50] Yuhong Nan, Xueqiang Wang, Luyi Xing, Xiaojing Liao, Ruoyu Wu, Jianliang Wu, Yifan Zhang, and XiaoFeng Wang. Are you spying on me? Large-scale analysis on IoT data exposure through companion apps. In *USENIX Security Symposium (USENIX Security)*, 2023.

[51] The International Association of Privacy Professionals (IAPP). US State Comprehensive Privacy Laws Report – Overview. https://iapp.org/resources/article/us-state-privacy-laws-overview/, 2024.

[52] Information Commissioner's Office. What is the legitimate interests basis? . https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/legitimate-interests/what-is-the-legitimate-interests-basis/, 2024.

[53] Shidong Pan, Dawen Zhang, Mark Staples, Zhenchang Xing, Jieshan Chen, Xiwei Xu, and Thong Hoang. Is it a trap? A large-scale empirical study and comprehensive assessment of online automated privacy policy generators for mobile apps. In *USENIX Security Symposium (USENIX Security)*, 2024.

[54] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. In *the Network and Distributed System Security Symposium (NDSS)*, 2019.

[55] Wenjun Qiu, David Lie, and Lisa Austin. Calpric: Inclusive and fine-grain labeling of privacy policies with crowdsourcing and active learning. In *USENIX Security Symposium (USENIX Security)*, 2023.

[56] Tamjid Al Rahat, Minjun Long, and Yuan Tian. Is your policy compliant? a deep learning-based empirical study of privacy policies' compliance with gdpr. In *the Workshop on Privacy in the Electronic Society (WPES)*, 2022.

[57] David Rodriguez, Ian Yang, Jose M Del Alamo, and Norman Sadeh. Large language models: A new approach for privacy policy analysis at scale. *Computing*, 106(12):3879–3903, 2024.

[58] Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson, and Norman M Sadeh. Automatic extraction of opt-out choices from privacy policies. In *AAAI Fall Symposia*, 2016.

[59] Mukund Srinath, Soundarya Sundareswara, Pranav Venkit, C. Lee Giles, and Shomir Wilson. Privacy lost and found: An investigation at scale of web privacy policy availability. In *ACM Symposium on Document Engineering (DocEng)*, 2023.

[60] Mukund Srinath, Shomir Wilson, and C Lee Giles. Privacy at scale: Introducing the privaseer corpus of web privacy policies. In *the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[61] Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, et al. PolicyGPT: Automated analysis of privacy policies with large language models. *arXiv preprint arXiv:2309.10238*, 2023.

[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[63] Rahmadi Trimananda, Hieu Le, Hao Cui, Janice Tran Ho, Anastasia Shuba, and Athina Markopoulou. OVRseen: Auditing network traffic and privacy policies in oculus VR. In *USENIX Security Symposium (USENIX Security)*, 2022.

[64] Isabel Wagner. Privacy policies across the ages: content of privacy policies 1996–2021. *ACM Transactions on Privacy and Security (TOPS)*, 26(3):1–32, 2023.

[65] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NIPS)*, 2022.

[66] Long Wen, Jinfei Liu, Feng Xue, Jian Lou, Zhibo Wang, Kui Ren, et al. CAPP-130: A corpus of Chinese application privacy policy summarization and interpretation. In *Advances in Neural Information Processing Systems (NIPS)*, 2023.

[67] 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. The creation and analysis of a website privacy policy corpus. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[68] Anhao Xiang, Weiping Pei, and Chuan Yue. PolicyChecker: Analyzing the GDPR completeness of mobile Apps' privacy policies. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2023.

[69] Yue Xiao, Zhengyi Li, Yue Qin, Xiaolong Bai, Jiale Guan, Xiaojing Liao, and Luyi Xing. Lalaine: Measuring and characterizing non-compliance of Apple privacy labels. In *USENIX Security Symposium (USENIX Security)*, 2023.

[70] Qinge Xie, Shujun Tang, Xiaofeng Zheng, Qingran Lin, Baojun Liu, Haixin Duan, and Frank Li. Building an open, robust, and stable voting-based domain top list. In *USENIX Security Symposium (USENIX Security)*, 2022.

[71] Fangyi Yu, Lee Quartey, and Frank Schilder. Legal prompting: Teaching a language model to think like a lawyer. In *the Workshop on Natural Legal Language Processing (NLLP)*, 2022.

[72] Razieh Nokhbeh Zaeem and K Suzanne Barber. The effect of the GDPR on privacy policies: Recent progress and future promise. *ACM Transactions on Management Information Systems (TMIS)*, 12(1):1–20, 2020.

[73] Kaifa Zhao, Le Yu, Shiyao Zhou, Jing Li, Xiapu Luo, Yat Fei Aemon Chiu, and Yutong Liu. A fine-grained Chinese software privacy policy dataset for sequence labeling and regulation compliant identification. In *the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[74] Lu Zhou, Chengyongxiao Wei, Tong Zhu, Guoxing Chen, Xiaokuan Zhang, Suguo Du, Hui Cao, and Haojin Zhu. POLICYCOMP: Counterpart comparison of privacy policies uncovers overbroad personal data collection practices. In *USENIX Security Symposium (USENIX Security)*, 2023.

[75] Sebastian Zimmeck and Steven M. Bellovin. Privee: An architecture for automatically analyzing web privacy policies. In *USENIX Security Symposium (USENIX Security)*, 2014.

[76] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. Maps: Scaling privacy compliance analysis to a million apps. In *Privacy Enhancing Technologies Symposium (PETS)*, 2019.

A Label Mapping for the OPP-115 Dataset

In the OPP-115 dataset, each data practice in a privacy policy segment is categorized into one of ten main categories, with further category-specific attributes. The ten main categories are: 1) *First-Party Collection/Use*, 2) *Third-Party Sharing/Collection*, 3) *User Choice/Control*, 4) *User Access, Edit and Deletion*, 5) *Data Retention*, 6) *Data Security*, 7) *Policy Change*, 8) *Do Not Track*, 9) *International and Specific Audiences*, and 10) *Other*.

We define a direct match between OPP-115 and our privacy policy clauses as a case where an OPP-115 main category sufficiently contains the information required by one of our privacy policy clauses. Since OPP-115 was annotated before the enactment of GDPR and US state privacy laws, only the *Data Retention* category aligns directly with clause P7. More specifically, a policy segment labeled *Data Retention* indicates that it contains a privacy practice specifying the retention period for collected user information, which is required by clause P7.

In addition to direct matches, some OPP-115 main categories are related to certain clauses but do not necessarily contain the required information. For example, a policy segment categorized under the main category *First-Party Use/-Collection* does not necessarily include information about the Categories of PI Collected, i.e., clause P1. Thus, we apply category-specific attributes as additional filters for indirect matches. We use the category-specific attribute *Personal Information Type* to match clauses P1 and P2, which represent the category of information collected by a business. OPP-115 classifies personal information into 14 distinct categories, along with two additional categories: *Other* and *Unspecified*. Policy segments labeled as *First-Party Collection/Use* where the attribute *Personal Information Type* is specified (i.e., not *Unspecified*) are used as positive samples for clause P1, and similarly, segments labeled as *Third-Party Sharing/Collection* are used for clause P2.

For the category of data purposes, we use the category-specific attribute *Purpose*, which OPP-115 classifies into 9 categories, along with *Other* and *Unspecified*. Similarly, segments labeled as the two main categories mentioned above (First-Party and Third-Party), where the *Purpose* attribute is specified, are used as positive samples for clauses P4 and P5. Finally, for third-party recipients, we use the segments labeled as *Third-Party Sharing/Collection* where the attribute *Third-Party Entity* is specified as positive samples. We also use these category-specific attributes in our evaluation of the inspection tasks, as discussed in Section 4.2.

B Prompt Templates

Figure 4 presents the prompt template for the coverage task, using clause P4 as the example. We set the system prompt as an initial instruction to provide background information on the tasks, which remains the same for both coverage and inspection tasks. The template also illustrates how we test one-shot and CoT prompting. For CoT prompting, we use the sentence: "*Let's work this out step by step to ensure we have the right answer*", to enable step-by-step inference by the LLM, similar to the prompt used in [43].

All prompts for the inspection tasks, as well as the coverage tasks for each clause, are available in our GitHub repository.

You are a knowledgeable, helpful, and honest assistant. You have deep expertise in current privacy regulations, including the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and other U.S. state privacy laws. You have strong skills in analyzing and explaining online privacy policies provided by businesses on their websites, ensuring clarity and accuracy in interpreting compliance requirements. **System**

Task Description and Instructions: \nI will provide a segment of a business's privacy policy from their website in Markdown format. Your task is to review the segment carefully to determine whether the segment includes the information about the purposes for the business to collect or use consumers' personal information. [Here is an example text: "…"] 1-shot

Focus: \nAnalyze exclusively for purpose information related to first-party collection/use of data. Ignore any purpose information related solely to third-party data sharing, disclosure, or selling.

CoT

Response Format: \n[Let's work this out in a step by step way to be sure we have the right answer. Finally,] Respond using the following JSON format: [<presence_flag>, <extracted_texts>].

- <presence_flag>: Set to '1' if the segment includes the purpose information for first-party data collection/use, or '0' if it does not.
- <extracted_texts>: A list containing the exact text of each purpose information for first-party data collection/use identified in the segment. If none is present, return 'None'.

The following text enclosed in double quotes is a segment of the privacy policy presented in Markdown format: \n"[Policy Segment]" **User**

Figure 4: Prompt template for the coverage task (using clause P4 as the example).

C Zero-shot and CoT

Few-shot learning and Chain-of-Thought (CoT) [40, 65] prompting techniques have been employed to enhance LLM inference. Here, we also compare the performance of coverage tasks using one-shot and CoT prompting. Table 11 lists the performance (F1-score) on coverage tasks for each clause using one-shot learning and CoT.

We find that CoT typically reduces task performance, primarily due to overfitting—where the model becomes overly specific to the clause while compromising the broader applicability of the task description. One-shot learning brings little to no improvement, indicating that the instructions used in our zero-shot prompts are sufficient for most cases.

Cl.	Dataset	1-shot	CoT
P1	OPP-115	0.85	0.83
P2	OPP-115	0.90	0.89
P3	Ours	0.87	0.83
P4	OPP-115	0.91	0.92
P5	OPP-115	0.93	0.93
P6	OPP-115	0.93	0.91
P7	OPP-115/PPGDPR	0.97/0.96	0.96/0.94
R1	PPGDPR/C3PA	0.91/0.96	0.90/0.94
R2	PPGDPR	0.94	0.96
R3	C3PA	0.97	0.97
R4	C3PA	0.98	0.98
R5	C3PA	0.93	0.92
R6	PPGDPR	0.86	0.91
R7	PPGDPR/C3PA	0.93/0.95	0.91/0.92
R8	PPGDPR	0.97	0.97
R9	Ours	0.91	0.90
R10	C3PA	0.99	0.99
E1	C3PA	0.93	0.93
E2	Ours	0.95	0.95
E3	PPGDPR	0.86	0.85
E4	Ours	0.98	0.98
E5	Ours	1.00	1.00
E6	Ours	0.96	0.94
E7	Ours	0.88	0.84
E8	Ours	0.99	0.99
D1	Ours	1.00	1.00
D2	Ours	1.00	1.00
D3	Ours	0.94	0.92
D4	Ours	0.98	0.98
D5	C3PA	0.99	0.99
D6	Ours	1.00	1.00
D7	Ours	1.00	1.00
D8	Ours	0.94	0.91
D9	Ours	0.92	0.90

Table 11: F1-scores from evaluating our method's performance across coverage tasks, using one-shot and CoT.