Analyzing Corporate Privacy Policies using AI Chatbots

Ziyuan Huang
University of Michigan
Ann Arbor, MI, USA
ziyuanh@umich.edu

Jiaming Tang
University of Michigan
Ann Arbor, MI, USA
jmtang@umich.edu

Manish Karir
SignetRisk Analytics, Inc.
Ann Arbor, MI, USA
mkarir@signetrisk.com

Mingyan Liu
University of Michigan
Ann Arbor, MI, USA
mingyan@umich.edu

Armin Sarabi
University of Michigan
Ann Arbor, MI, USA
arsarabi@umich.edu

Abstract

In this paper, we present and evaluate an automated pipeline for the large-scale analysis of corporate privacy policies. Organizations usually develop their privacy policies in isolation to best balance their business needs, user rights, as well as regulatory requirements. A wide-ranging and structured analysis of corporate privacy policies is essential to facilitate a deeper understanding of how organizations have balanced competing requirements. Our approach consists of a web crawler that can navigate to and scrape content from web pages that contain privacy policies, and a set of AI chatbot task prompts to process and extract structured/labeled annotations from the raw data. The analysis includes the types of collected user data, the purposes for which data is collected and processed, data retention and protection practices, and user rights and choices. Our validation shows that our annotations are highly accurate and consistent. We use this architecture to gather data on the privacy policies of companies in the Russell 3000 index, resulting in hundreds of thousands of annotations across all categories. Analysis of the resulting data allows us to obtain unique insights into the state of the privacy policy ecosystem as a whole.

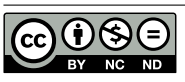
CCS Concepts

• **Social and professional topics** → **Privacy policies**; • **Computing methodologies** → **Information extraction**; • **Information systems** → *Web crawling*.

Keywords

Privacy policies, AI chatbots, large language models, text annotation, web crawling

ACM Reference Format:
Ziyuan Huang, Jiaming Tang, Manish Karir, Mingyan Liu, and Armin Sarabi. 2024. Analyzing Corporate Privacy Policies using AI Chatbots. In *Proceedings of the 2024 ACM Internet Measurement Conference (IMC '24), November 4–6, 2024, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3646547.3689015



IMC '24, November 4–6, 2024, Madrid, Spain

ACM ISBN 979-8-4007-0592-2/24/11
https://doi.org/10.1145/3646547.3689015

1 Introduction

Privacy policies are important legal documents that specify the rights and responsibilities of both the corporate entity as well as their clients. As important as they are, these documents lack a standard format, are often lengthy with legalese, and all in all are not written to encourage reading for an ordinary person. Perhaps by design, users seldom understand corporations' data collection practices or their own rights in this regard. Previous attempts to analyze these policies have typically been small scale as they require intensive manual labor in parsing and annotation[12, 17, 20].

The emergence of large language model (LLM) based AI chatbots presents a unique opportunity to create automated data processing pipelines to analyze these natural language documents. With appropriate tasking, it is possible to craft an automated analysis of a given policy document that cuts through the dense legal text and distills out key relevant structured data elements. Care must be taken to ensure the validity of the resulting data. Our automated data pipeline depicted in Figure 1 can generate human- and machine-readable summaries (annotations) of privacy policies. Validation results show that these annotations are highly accurate, consistent, and significantly more detailed (for collected data types and data collection purposes) than previous work. The resulting structured and consistent data is ideal for a wide-ranging analysis of the privacy policy ecosystem as it exists today. Our main contributions are as follows:

(1) A unique flexible/programmable data pipeline for the analysis of privacy policies including manual taxonomy construction and automated corpus annotation, supporting out-of-vocabulary (zero-shot) annotations by leaving the set of labels open (e.g., by capturing/categorizing terms not explicitly mentioned in our prompts);
(2) More fine-grained and consistent annotations than prior work via a comprehensive and extendable taxonomy;
(3) A new large-scale structured privacy policy dataset with fine-grained and consistent annotations which enables a wide range of new statistical, risk and legal analysis of the privacy policy landscape.[1]

Our own analysis of the privacy policies of companies in the Russell 3000 highlights some interesting high level insights such as

[1] Our dataset, named AIPAN-3k (AI-driven Privacy policy ANnotations of Russell 3000), can be accessed at https://github.com/arsarabi/aipan-3k.

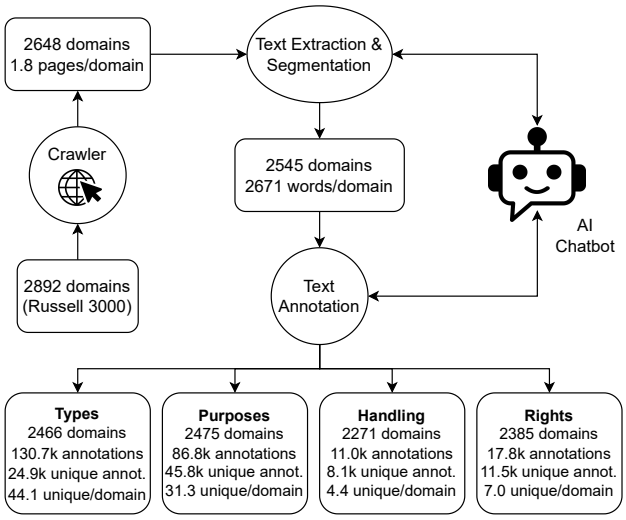


Figure 1: Overview of our pipeline. We leverage an AI chatbot to process crawled privacy policies from company websites and produce labeled annotations.

the ubiquitous use of data collection for basic operations and enhancing user experience, the lack of opt-in versus opt-out policies, vagueness in the duration of data retention, the lack of explicitly stated data protections, and finally the extreme reliance of companies in the consumer discretionary sector on broad data collections. While people have often highlighted failings in individual privacy policies, it is only through large-scale analysis that the full scale of specific structural issues can be brought to light.

2 Related Work

Several prior studies have focused on developing datasets related to privacy policies, including datasets of crawled website privacy polices [1] and human-annotated datasets examining data collection and privacy practices [17, 20], opt-out choices [5], and those designed for question answering [12]. Other studies have tackled automated analysis of privacy policies using, e.g., semantic ontology reasoning [2, 3, 10], traditional ML models (including naive Bayes, LDA, SVM, etc.) [9, 13, 19, 21], and deep learning models [7, 8, 18]. With the widespread use of language models, researchers have started to use LLMs to analyze privacy policies, e.g., to enable question answering [12], to detect text segments discussing certain data privacy aspects [15], or to process mobile app privacy policies [11]. To the best of our knowledge, ours is the first work that leverages LLM-based chatbots for the *annotation* of privacy policies based on a flexible taxonomy to create structured and normalized summaries of data privacy practices. This allows our framework to be easily extended through continuous improvement of our prompts and the use of annotated data to train downstream ML models.

Taxonomy frequently serves as the first step towards formalizing analytical scenarios such as policy making, legislation, and statistical analysis. Solove [14] first categorizes privacy policies by the social understanding of privacy violations. More recent efforts [4, 6, 17, 18] have emerged from the rising demands of comprehending and securing data in the digital world. All of these primarily involve various levels of human annotations. This limits flexibility due to the annotators' degrading performance on large tasks. Our emphasis on automation and consistency by leveraging AI chatbots allows us to implement taxonomy annotation reliably and repeatably, allowing our approach to be scaled as needed.

3 Data Collection and Processing

In this section, we describe our data collection and processing pipeline, which includes a web crawler for scraping website privacy policies and a set of AI chatbot prompts for extracting labeled annotations from privacy policy texts.

3.1 Acquisition of target privacy policies

We use the constituents of the Vanguard Russell 3000 ETF as of March 31, 2024, resulting in 2916 company names alongside their corresponding S&P (Standard and Poor) sectors (comprised of 11 different sectors). To find each company's Internet domain, we retrieve the first Google search result for the associated company name. We then manually review the results for accuracy, which yields 2892 unique domains for our study. The number of domains is smaller than the initial number of companies due to duplicates, e.g., GOOGL and GOOG both belonging to Alphabet Inc.

We then develop a web crawler using the Crawlee library (employing the Playwright backend with a headless Chromium browser)[2] to navigate to and scrape privacy policies from websites. To find potential *privacy pages*, i.e., web pages containing a privacy policy, we follow up to three links containing the word "privacy" from the bottom of a website's homepage (also used by prior work [1]), try to navigate to `/privacy-policy` and `/privacy`,[3] and follow up to five links containing the word "privacy" from the top of the aforementioned five pages (i.e., `/privacy-policy`, `/privacy`, and up to three pages linked to from the homepage). Note that the latter allows us to find privacy policies for websites with a dedicated privacy home/center page, with the actual privacy policy found by following an additional link. This leads to a maximum of 31 pages crawled from each website, though the average number of crawled pages (including the homepage) is 5.1. Of our 2892 domains, our crawler managed to successfully (i.e., resulting in an HTTP status code below 400) navigate to at least one potential privacy page for 2648 (91.6%) domains. We then remove duplicates and non-English pages, yielding an average of 1.8 potential privacy pages per successful domain. We provide an analysis of failed crawls, among other failures, in section 4.

3.2 Annotation using AI chatbots

We now describe how we extract relevant text content from our crawled data, and produce structured annotations summarizing the company's data privacy practices. To this end, we design a set of task prompts for an AI chatbot to split scraped content into sections, and then extract and label mentions of collected data types, data collection purposes, data retention and protection practices,

[2] https://crawlee.dev, https://playwright.dev

[3] `/privacy-policy` and `/privacy` point to an existing page for 1577 (54.5%) and 1405 (48.6%) of our domains, respectively

and user rights and choices. We create detailed instructions for each task, including an input/output example. We further refine prompts through an iterative process by examining their outputs for example inputs and then revising instructions to address common errors. This results in highly accurate outputs as discussed in section 4. We have included examples of our prompts in Figure 2 in Appendix C. Our pipeline is evaluated using OpenAI's `gpt-4-turbo-2024-04-09` model.

3.2.1 Dividing into sections. To extract relevant text from our crawled web pages, we first convert the page's HTML into text using the inscriptis library [16], and then divide it into sections discussing the following aspects of a privacy policy (based on aspects from Wilson et al. [17]):

- `Types`: What types or categories of data are collected.
- `Methods`: How data may be collected, including methods, sources, or tools used for data collection.
- `Purposes`: What are the purposes of data collection, including why data is collected and how it is used.
- `Handling`: How the collected data is handled, stored, or protected, including data processing, data retention, and security mechanisms.
- `Sharing`: Whether and how data is shared with or disclosed to third parties.
- `Rights`: User rights, choices, and controls, including access, edit, deletion, and opt-out options.
- `Audiences`: Information related to specific audiences, e.g., children or users from California, Europe, etc.
- `Changes`: If and how users will be informed of changes.
- `Other`: Information not covered above, including introductory or generic statements, contact information, and other information not directly related to data privacy.

Dividing the text into sections helps remove unrelated content and minimize token usage for subsequent annotation tasks. To do this, we follow the two-step process detailed in Appendix B, consisting of (1) dividing into sections by means of detecting and labeling section headings according to the above aspects, and (2) analyzing the entire text content to divide it into sections if the previous approach fails. This results in successful extractions for 2545 of our domains (88% of all domains, and 96.1% of domains with a successful crawl). We define a successful extraction as being able to extract text corresponding to any aspects other than `audiences`, `changes`, or `other`. We ignore `audiences` since some websites have dedicated pages for specific jurisdictions, which is not the focus of this study. The median length of a privacy policy (excluding `audiences`, `changes`, and `other`) is 2671 words.

3.2.2 Annotating data privacy practices using context. To generate machine-readable annotations that summarize practices covered in a privacy policy, we first feed the corresponding section (`types`, `purposes`, `handling`, or `rights`) to the chatbot, falling back to feeding the entire text if the former does not produce any annotations.[4] The latter helps increase coverage if the privacy policy does not have a dedicated section for the associated aspect, e.g., short policies or those covering an aspect in line with other practices. This approach helps improve accuracy by ensuring that only relevant text sections are provided to the chatbot for annotation. We generate annotations for each aspect as described below. Figure 1 includes a summary of our annotations, with more details provided in Table 1. Note that to detect and remove hallucinations, we programmatically verify that the chatbot-generated annotations are indeed present in the privacy policy text.

Data types: We first create a chatbot task for extracting relevant mentions verbatim from the text, and feed extracted text into another task aimed at categorizing data types and generating *normalized* descriptors (e.g., mapping both "mailing address" and "home address" to "postal address" and categorizing them as "Contact info"). To do this, we examine annotations from the first task and logically divide them into 6 meta-categories, 34 categories, and a non-exhaustive list of 125 normalized descriptors, resulting in a much more granular taxonomy than prior work [17]. These are then compiled into a glossary and attached to both prompts; this helps provide the chatbot with more context for performing the tasks. Finally, we ask the chatbot to generate descriptors of its own for data types not listed in the glossary. Table 1 includes a subset of our categories and their associated counts (the full version is included in Table 4 in Appendix D), as well as the three most common descriptors for each category.

Data collection purposes: To extract specific purposes for data collection, we use a similar approach by asking the chatbot to extract relevant mentions, and normalize them according to a manually curated glossary with 7 categories and 48 descriptors. Our results are summarized in Table 1.

Data handling and user rights: We also annotate mentions of data retention periods, specific data protection measures, opt-in/opt-out choices and privacy settings, and users' access to view, edit or delete their data. This is achieved using two chatbot tasks (one for data retention/protection and one for user choices/access) that extract relevant mentions and label them according to a set of labels based on practices defined by Wilson et al. [17] and included in Table 1.

4 Validation

To validate the performance of our pipeline we first examined domains with unsuccessful crawls (244 domains) or text extraction (103 domains). We manually examined 50 randomly selected failures and found 27 not containing a privacy policy, 11 crawler-related failures (6 crawler exceptions/timeouts, 3 blocked crawls, and 2 failures due to dynamic JavaScript-loaded content), 5 failures to detect relevant links (3 links not containing the word privacy, e.g., "Legal Notices", one link triggering a JavaScript action, and one link in the website's consent box and not captured by Playwright), 5 privacy policies in PDF format (which is currently not supported by our pipeline), and 2 non-English websites.

Of the remaining 2545 domains with a successful privacy policy extraction, 2529 received at least one annotation, and 375 did not receive any annotations for at least one of `types`, `purposes`, `handling`, and `rights`. We manually inspected 20 such domains and found that 16 of them indeed did not include details regarding the missing aspects. Of the remaining four, our crawler did not extract parts of the privacy policy for three (one due to dynamically

[4] This fallback is activated at least once for 708/2545 privacy policies.

Table 1: Summary of AI-generated annotations, including collected data types, data collection purposes, data retention/protection practices, and user choices/access. The reported counts correspond to the number of unique annotations after eliminating repetitive mentions of the same term for each privacy policy. For data types and collection purposes, we also report the top 3 descriptors and their frequency within each category; for data handling and user rights, we include a brief description of the practice. For data types, we only show the top 4 most mentioned categories; the full version can be found in Table 4 in Appendix D.

	Meta-category	Category	Description(s)
Types (108,748)	Physical profile (36,158)	Contact info (10,582)	email address (27.3%), postal address (25.6%), phone number (25.1%)
		Personal identifier (9,534)	name (31.0%), unique personal identifier (11.7%), social security number (8.6%)
		Professional info (7,779)	employment history (16.3%), employer details (10.8%), job title (10.5%)
		Demographic info (6,203)	gender (14.1%), age (10.6%), demographic info (9.9%)
	Digital profile (19,182)	Device info (8,659)	browser type (22.4%), operating system (15.6%), device identifier (12.9%)
		Online identifier (4,283)	ip address (65.5%), online identifier (9.1%), domain name (3.9%)
		Account info (3,403)	username (30.1%), password (19.1%), account info (9.0%)
		Network connectivity (1,191)	isp (21.6%), internet connection (17.3%), network traffic (8.0%)
	Bio/health profile (4,751)	Medical info (2,929)	medical info (14.7%), medical conditions (10.1%), disability status (4.3%)
		Biometric data (1,187)	biometric data (25.0%), facial data (12.6%), fingerprint (10.9%)
		Physical characteristic (427)	physical characteristics (46.6%), weight (7.3%), height (6.3%)
		Fitness & health (208)	physical activity info (25.0%), sleep patterns (17.3%), health metrics (3.8%)
	Financial/legal profile (8,864)	Financial info (4,955)	payment card info (25.6%), financial info (15.3%), bank account info (14.7%)
		Legal info (1,729)	signature (21.2%), background checks (9.8%), criminal records (7.2%)
		Financial capability (1,399)	income (17.6%), credit history (13.9%), credit score (7.6%)
		Insurance info (781)	health insurance (29.2%), insurance policy number (19.5%), insurance info (9.7%)
	Physical behavior (4,375)	Precise location (2,389)	gps location (54.8%), precise location (13.0%), device location (4.1%)
		Approximate location (1,620)	country (18.7%), zip code (18.0%), approximate location (17.6%)
		Travel data (276)	movement patterns (26.1%), travel history (10.9%), travel data (2.2%)
		Physical interaction (90)	in-store interactions (43.3%), event participation (4.4%), interactions (4.4%)
	Digital behavior (26,975)	Internet usage (7,847)	browsing history (14.5%), search history (8.3%), click behavior (7.7%)
		Tracking data (3,486)	cookies (43.4%), web beacons (19.0%), online tracking technologies (6.8%)
		Product/service usage (3,076)	user engagement metrics (20.6%), website usage (9.7%), app usage (9.1%)
		Transaction info (2,721)	purchase history (28.6%), transaction info (9.5%), commercial info (5.5%)
Purposes (77,360)	Operations (47,997)	Basic functioning (27,564)	cust. service (9.3%), cust. communication (8.0%), transaction processing (4.8%)
		User experience (10,603)	product improvement (20.1%), personalization (16.3%), quality assurance (4.4%)
		Analytics & research (9,830)	analytics (17.4%), product/service development (8.6%), research (6.2%)
	Legal (19,086)	Legal & compliance (10,142)	legal compliance (28.1%), regulatory compliance (10.2%), policy compliance (7.4%)
		Security (8,944)	fraud prevention (21.8%), authentication (6.6%), product/service safety (5.4%)
	Third-party (9,694)	Advertising & sales (8,107)	direct marketing (20.8%), promotions (18.8%), targeted advertising (16.3%)
		Data sharing (1,587)	third-party sharing (18.8%), sharing with partners (15.0%), anonymization (4.3%)
Handling (10,014)	Data retention (4,550)	Limited (3,843)	Retention period is limited but unspecified.
		Stated (555)	Retention period is specified (and extracted by the chatbot).
		Indefinitely (152)	Collected data is retained indefinitely.
	Data protection (5,464)	Generic (3,076)	Generic statement regarding data protection/security.
		Access limit (646)	Data access is restricted on a need-to-know basis.
		Secure transfer (459)	Data transfer is secured, e.g., via encryption.
		Secure storage (490)	Data is stored securely, e.g., in an encrypted format or database.
		Privacy program (413)	Company has a data privacy/protection program.
		Privacy review (238)	Privacy measures and data protection practices are reviewed/audited.
		Secure authentication (142)	User authentication is secured, e.g., via encryption or 2FA.
Rights (16,605)	User choices (7,484)	Opt-out via contact (3,976)	Users must directly contact the company (e.g., via email) to opt-out.
		Opt-out via link (1,915)	Users can opt-out via a link provided by the company.
		Privacy settings (728)	Company provides controls via a dedicated privacy settings page.
		Opt-in (550)	Users must consent before data can be collected, used, or shared.
		Do not use (315)	The only option is for users to not use a feature or service.
	User access (9,121)	Edit (3,591)	Users can modify, correct, or delete specific data.
		Full delete (1,948)	Users can fully delete their account (all data is removed from servers/databases).
		View (1,680)	Users can view their data.
		Export (1,499)	Users can export or obtain a copy of their data.
		Partial delete (336)	Users can partially delete their account (company may retain some of their data).
		Deactivate (67)	Users can deactivate their account (company retains access to their data).

Table 2: Breakdown of collected data types (top) and data collection purposes (bottom). For each (meta-)category we report the overall coverage (percentage of companies with at least one relevant annotation), and the average and standard deviation of the number of unique mentions (descriptors). We also provide a sector breakdown by reporting statistics on the top 3 sectors with the highest within-sector coverage, and the sector with the lowest. A breakdown of data types over all categories is provided in Table 5 in Appendix D. We use the following S&P sector abbreviations: CD: Consumer discretionary, CS: Consumer staples, EN: Energy, FS: Financials, HC: Health care, IN: Industrials, IT: Information technology, MT: Materials, RE: Real estate, TC: Communication services, UT: Utilities.

(a) Collected data types (see Table 4 in Appendix D for full version).

Meta-category	Overall statistics		Sector statistics (sorted by coverage)											
	Coverage	Mean/SD	Highest			2nd highest			3rd highest			Lowest		
Physical profile	**92.6%**	**12.8±11.5**	TC	94.9%	13.2±9.4	HC	94.5%	12.8±11.1	IT	93.8%	12.9±11.3	MT	86.0%	10.0±9.4
Digital profile	87.1%	7.5±5.4	TC	94.9%	10.0±6.7	UT	94.4%	6.2±4.9	CD	92.8%	9.5±6.2	FS	74.4%	7.4±5.2
Bio/health profile	**34.5%**	**5.0±5.4**	HC	51.7%	5.9±5.8	**CD**	**39.5%**	**4.2±4.8**	FS	35.2%	6.1±6.7	**EN**	**12.1%**	**2.8±2.0**
Financial/legal profile	60.7%	5.2±4.9	FS	78.3%	7.1±6.6	**CD**	**76.6%**	**4.9±4.1**	TC	69.4%	4.1±3.6	**EN**	**33.3%**	**4.8±3.9**
Physical behavior	62.5%	2.4±1.8	TC	80.6%	3.1±2.3	**CD**	**76.6%**	**2.9±2.1**	CS	68.9%	2.3±1.3	**EN**	**37.4%**	**1.9±1.6**
Digital behavior	90.1%	10.3±8.3	TC	96.9%	12.9±8.8	**CD**	**94.8%**	**14.1±9.8**	IT	92.9%	11.8±9.0	MT	80.7%	9.0±6.7

(b) Data collection purposes.

(Meta-)category	Overall statistics		Sector statistics (sorted by coverage)											
	Coverage	Mean/SD	Highest			2nd highest			3rd highest			Lowest		
Operations	**97.5%**	**15.6±11.8**	TC	100.0%	17.3±12.4	CS	100.0%	17.7±12.1	UT	100.0%	14.2±11.4	**EN**	**92.9%**	**10.3±9.5**
- Basic functioning	95.1%	9.1±7.8	CS	99.0%	9.7±8.5	TC	98.0%	8.7±7.7	HC	97.4%	8.9±7.7	**EN**	**88.9%**	**6.1±5.7**
- User experience	86.5%	3.9±2.9	CS	93.2%	4.7±3.4	IT	92.3%	4.1±3.1	CD	92.1%	4.4±2.9	FS	75.1%	3.5±2.5
- Analytics & research	81.3%	4.1±3.1	**CD**	**89.3%**	**4.3±3.0**	TC	88.8%	5.0±3.4	CS	87.4%	4.3±2.8	**EN**	**66.7%**	**3.0±2.5**
Legal	82.0%	7.3±5.9	TC	89.8%	6.9±4.7	CD	86.6%	7.9±6.1	FS	85.2%	7.3±6.3	**EN**	**62.6%**	**5.4±5.1**
- Legal & compliance	73.2%	4.1±3.3	TC	82.7%	3.5±2.5	FS	78.3%	4.1±3.2	CD	78.0%	4.1±3.2	**EN**	**47.5%**	**3.5±2.5**
- Security	72.5%	4.1±3.3	TC	85.7%	3.9±2.9	CS	79.6%	3.9±2.7	CD	79.0%	4.6±3.6	**EN**	**53.5%**	**3.3±3.4**
Third-party	81.2%	3.5±2.9	CD	91.4%	4.2±2.8	CS	87.4%	4.0±2.6	IT	85.8%	3.8±2.6	**EN**	**61.6%**	**2.6±2.5**
- Advertising & sales	78.0%	3.0±2.3	**CD**	**91.1%**	**3.6±2.6**	CS	85.4%	3.6±2.5	IT	84.8%	3.3±2.1	**EN**	**51.5%**	**2.4±2.0**
- Data sharing	**26.1%**	**2.1±2.3**	TC	36.7%	2.0±1.2	RE	35.5%	1.7±1.2	HC	30.3%	2.8±4.0	FS	18.2%	1.8±1.6

Table 3: Data handling and user rights annotations.

Meta-category	Category	Cov.	Sector statistics		
			Highest	2nd highest	Lowest
Data retention	Limited	**60.9%**	TC 81.6%	IT 81.4%	UT 25.9%
	Stated	**9.9%**	IT 16.4%	TC 15.3%	UT 5.6%
	Indefinitely	5.5%	HC 6.5%	TC 6.1%	CD 4.5%
Data protection	Generic	**73.1%**	RE 78.2%	IT 76.5%	EN 63.6%
	Access limit	19.1%	FS 29.4%	IT 22.0%	MT 11.4%
	Secure transfer	**14.0%**	UT 18.5%	TC 18.4%	EN 7.1%
	Secure storage	**16.1%**	FS 31.6%	IT 21.4%	CS 4.9%
	Privacy program	9.9%	IT 16.4%	FS 14.3%	RE 3.2%
	Privacy review	6.8%	IT 13.0%	UT 11.1%	CS 2.9%
	Secure auth.	4.2%	FS 7.2%	IT 5.3%	MT 1.8%
User choices	Opt-out (contact)	**65.2%**	**TC 72.4%**	**IT 71.8%**	EN 43.4%
	Opt-out (link)	36.1%	TC 61.2%	CS 60.2%	EN 17.2%
	Privacy settings	**17.7%**	**TC** 29.6%	**IT 24.5%**	EN 8.1%
	Opt-in	**17.7%**	CS 22.3%	UT 22.2%	TC 12.2%
	Do not use	10.5%	UT 14.8%	CS 13.6%	RE 8.1%
User access	Edit	71.6%	**IT 85.4%**	**TC 80.6%**	EN 43.4%
	Full delete	**53.5%**	CD 63.9%	TC 62.2%	UT 27.8%
	View	45.6%	IT 57.3%	TC 52.0%	UT 27.8%
	Export	42.9%	IT 61.0%	CS 49.5%	UT 18.5%
	Partial delete	11.2%	TC 22.4%	IT 14.6%	UT 1.9%
	Deactivate	2.5%	TC 8.2%	UT 5.6%	IN 0.8%

loaded content, one due to content under an expandable HTML element, and one due to most of the privacy policy included as an image), and one combined privacy policies in different languages which led to it being discarded by our pre-processing step.

To examine the precision of our annotations, we manually inspected 340 annotations of collected data types (10 annotations per category), 175 data collection purposes (25 per category), 200 data retention/protection practices (10 per category), and 220 user choices and access (20 per category). The evaluation tasks were evenly divided among the authors of this paper. The evaluation of each author was further reviewed by another randomly selected author. The resulting estimated precision scores are 89.7% for data types, 94.3% for collection purposes, 97.5% for data handling and 90.5% for user rights.[5] Table 6 in Appendix D provides some examples of these annotations and the associated contextual text.

5 Data Analysis

We next discuss a number of statistical results on these privacy policies. Unless otherwise stated, percentages in the remainder of this section are computed over the 2529 companies that have at least one annotation.

A detailed analysis of our annotations of selected data types and collection purposes is provided in Table 2. We first report the overall coverage, defined as the percentage of companies with at

[5]Note that ~40% of errors for user rights belong to the "do not use" category which has proven particularly difficult to annotate accurately.

least one annotation in each of our categories. For companies that are counted toward the coverage (i.e., with at least one mention of, e.g., "Physical profile"), we then report the average and standard deviation of the number of unique mentions under the associated category, after removing repetitive mentions that are mapped by the chatbot to the same descriptor in Table 1. We provide a similar analysis of data handling and user rights in Table 3, where we report coverage levels for different practices. For both tables we also report statistics for the S&P sectors with the highest/lowest within-sector coverages. We highlight some of the more prominent/interesting findings from Tables 2 and 3 below.

Data types: We observe that 2365 (93.5%) of our companies collect data from at least 3 or more categories, with 1335 (52.8%) collecting more than 13, 329 (13.0%) collecting more than 22, and 122 (4.8%) collecting more than 25. A number of notable highlights from Table 2a are listed below.

- The vast majority of policies mention the collection of data on physical profiles of individuals, while the least mentioned category was health-related data (to be expected given the sensitive nature of health data).
- Consumer discretionary is the second highest sector in mentions of the health data category (somewhat surprisingly); it is also overall the most actively observed sector for any data category, with an average of 16.3 categories (48.8 distinct data descriptors) collected. When combined with observations in Table 2b, we note the primary use of these data is advertising/sales and analytics/research.
- The energy sector has the least number of mentions of data collection activities. However, and surprisingly, 12.1% of this sector collect health-related data and 33% mention the collection of financial/legal profile related data.

Collection purposes: The data summarized in Table 2b shows that almost all policies (97.5%) call out the use of collected data for monitoring or improving the services provided by the company. On the other hand, very few (26%) explicitly mention that data might be shared with third parties. Interestingly, an inspection of our granular descriptors reveals that 26 companies mention that collected data may be sold to third-parties (categorized as "data sharing → data for sale"). Similar to data types, those in the energy sector were observed to have the least mentions of the purposes of their data collection.

Data handling: Looking at Table 3, while over 60% of the policies mention that data is retained for a limited period of time, only 10% explicitly state a retention period. Similarly, over 70% provide a generic mention of data protection, but only 39.9% mention any specifics about the adoption of data protection practices (e.g., access limit, secure transfer/storage, and so forth).

- The median stated retention period is 2 years, with a minimum of 1 day (for "arescre.com" and "pg.com") and a maximum of 50 years (for "bms.com").
- Only 16.1% mention the use of secure storage; 14.0% mention the use of secure transfer of data in transit.

User rights: Table 3 shows that the ability to opt out (two-thirds) is far more common than opt-in options (<20%).

- Just over half mention the ability for users to request full deletion of their data (surprisingly).
- 77.5% provide read/write access (edit, partial delete, or full delete); 0.5% provide read-only access (only view/export).
- 22.0% do not have any mention of user access.

Not surprisingly, companies in the information technology and communication services sectors were most prominent in mentioning the ability of users to edit the collected data, opt-out of data collection, or provide detailed privacy settings.

6 Discussion and Conclusion

Large scale analysis of privacy policies is a challenging task since they are first and foremost legal documents. Using the architecture presented in this paper, we have shown it is possible to map these complex documents to well structured and normalized annotations. This crucial first step then unlocks the ability to perform a variety of statistical analyses such as trends, policy peer group comparisons, policy quality evaluations, as well as legal exposure risk analysis.

Our validation shows that our architecture produces highly accurate and consistent annotations, which can effectively mitigate the subjectivity in manual processing, arguably one of the main challenges in annotating this type of documents. We believe this method can be the first step toward establishing a standardized approach to analyzing privacy policies. While we have exclusively used GPT-4 Turbo for our study, in principle any LLM can be used for this purpose. To further justfy our choice, we conducted a comparison study of 20 randomly selected privacy policies using GPT-3.5 Turbo and Llama-3.1. GPT-3.5 exhibits an unsatisfactory performance as it seemingly struggles to understand the complex nature of privacy policy texts (e.g., mistaking the marketing platform ActiveCampaign as a data type describing campaign engagement), while Llama-3.1 achieves more comparable performance to GPT-4. We further manually validated the extractions for collected data types from GPT-4 and Llama-3.1 for the 20 privacy policies, observing that the former achieves a higher precision score (96.2%) than the latter (83.2%). Our experiments show that Llama-3.1 is unable to follow instructions as closely as GPT-4, e.g., it tends to extract data types mentioned in negated contexts (e.g., data mentioned after "this privacy notice does not apply to" in the privacy policy of Brown & Brown Insurance), even though we explicitly instruct the Chatbot to ignore such cases.

Our ongoing work continues to improve the fidelity of our annotations through continuous refinement of our taxonomy and providing the chatbot with clear instructions. For instance, our manual inspections reveal that mentions of unlimited retention periods often concern anonymized or aggregated data, which is less concerning than personally identifiable information. Providing the chatbot with instructions to ignore such mentions, or better yet, to extract the associated data types for various data privacy practices, can further increase the quality of our annotations. Finally, training offline LLMs to replicate the chatbot-generated annotations is another important aspect of our future work.

7 Acknowledgments

We would like to thank Alexa Faulkner, our shepherd, and the anonymous reviewers for their contributions and valuable feedback. This work is supported by the NSF under grant CNS-2012001.

A Ethics

Our primary approach towards raw data collection involves the use of web crawling as the underlying data collection technique. This technique itself has been widely used by Internet researchers over the past few decades. Though we take care to limit our use of this technique to only gather the minimum data necessary, it is possible that some organizations may object to our data collection activity. Our research purpose is limited solely to the examination of the publicly posted privacy policy by a company and there is no other technique for obtaining this information at the scale that is necessary for our research. Our use of this technique is therefore tempered, and necessary as the only approach for achieving the scale needed for our analysis.

References

[1] Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. Privacy Policies over Time: Curation and Analysis of a Million-document Dataset. In *Web Conference*. 2165–2176.

[2] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In *USENIX Security Symposium*. 585–602.

[3] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions Speak Louder than Words: Entity-sensitive Privacy Policy and Data Flow Analysis with PoliCheck. In *USENIX Security Symposium*. 985–1002.

[4] Annie I Antón, Julia Brande Earp, and Angela Reese. 2002. Analyzing Website Privacy Requirements Using a Privacy Goal Taxonomy. In *Joint International Conference on Requirements Engineering*. IEEE, 23–31.

[5] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, et al. 2020. Finding a Choice in a Haystack: Automatic Extraction of Opt-out Statements from Privacy Policy Text. In *Web Conference*. 1943–1954.

[6] Jose M Del Alamo, Danny S Guaman, Boni García, and Ana Diez. 2022. A Systematic Mapping Study on Automated Analysis of Privacy Policies. *Computing* 104, 9 (2022), 2053–2076.

[7] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *USENIX Security Symposium*. 531–548.

[8] Logan Lebanoff and Fei Liu. 2018. Automatic Detection of Vague Words and Sentences in Privacy Policies. *arXiv preprint arXiv:1808.06219* (2018).

[9] Aaron K Massey, Jacob Eisenstein, Annie I Antón, and Peter P Swire. 2013. Automated Text Mining for Requirements Analysis of Policy Documents. In *International Requirements Engineering Conference*. IEEE, 4–13.

[10] Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B Norton, N Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. 2018. PrivOnto: A Semantic Framework for the Analysis of Privacy Policies. *Semantic Web* 9, 2 (2018), 185–203.

[11] Shidong Pan, Thong Hoang, Dawen Zhang, Zhenchang Xing, Xiwei Xu, Qinghua Lu, and Mark Staples. 2023. Toward the Cure of Privacy Policy Reading Phobia: Automated Generation of Privacy Nutrition Labels from Privacy Policies. *arXiv preprint arXiv:2306.10923* (2023).

[12] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 4949–4959.

[13] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the Provision of Choices in Privacy Policy Text. In *Conference on Empirical Methods in Natural Language Processing*. 2774–2779.

[14] Daniel J Solove. 2006. A Taxonomy of Privacy. *University of Pennsylvania Law Review* 154, 3 (2006), 477–564.

[15] Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, et al. 2023. PolicyGPT: Automated Analysis of Privacy Policies with Large Language Models. *arXiv preprint arXiv:2309.10238* (2023).

[16] Albert Weichselbraun. 2021. Inscriptis - A Python-based HTML to Text Conversion Library Optimized for Knowledge Extraction from the Web. *Journal of Open Source Software* 6, 66 (2021), 3557.

[17] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1330–1340.

[18] Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, et al. 2018. Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations. *Transactions on the Web* 13, 1 (2018), 1–29.

[19] Sebastian Zimmeck and Steven M Bellovin. 2014. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In *USENIX Security Symposium*. 1–16.

[20] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. *Privacy Enhancing Technologies* (2019).

[21] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven Bellovin, and Joel Reidenberg. 2016. Automated Analysis of Privacy Requirements for Mobile Apps. In *AAAI Fall Symposium Series*.

B Process for dividing privacy policies into sections

We refer to the process of dividing a privacy policy document according to the aspects defined in subsection 3.2.1 as *segmentation*. The two primary steps for segmentation are described below.

Segmentation based on headings: We first detect headings by extracting text wrapped in HTML heading tags (`<h1>` through `<h6>`), as well as bold text (`<b>` and `<strong>`) that appears on a separate line (and not inline with non-bold text). If a page contains more than five headings, we then divide it into sections by assigning each piece of text to the first heading preceding it. We then generate a table of contents for the page while recognizing the section hierarchy implied by heading levels (i.e., `<h1>` through `<h6>` followed by bold text). Finally, we create a chatbot task that takes in a table of contents and labels them according to the nine aspects defined in subsection 3.2.1.

Segmentation via text analysis: The previous process may fail to extract relevant text due to insufficient or non-descriptive headings, e.g., if headings do not use appropriate HTML tags, or for short policies without any headings. To deal with these malformatted/short polices, we first detect domains that do not yield any text for at least one of `types`, `purposes`, `handling`, and `rights` (which are the main focus of this study). We then process these domains by feeding pages' entire text into a chatbot tasked with both dividing it into sections and labeling those sections accordingly.

C Example prompts

Task: Assume the role of a data privacy expert tasked with analyzing website privacy policies. Use the provided glossary to label a list of section headings according to the categories given below:

- **types**: What types or categories of data are collected.
- **methods**: How data may be collected, including methods, sources, or tools used for data collection.
- **purposes**: What are the purposes of data collection, including why data is collected and how it is used.
- …

Carefully follow the instructions below, using the provided glossary and example as a guide.

Instructions:

(1) Carefully and thoroughly read the section headings (extracted from text that may contain a privacy policy) provided in the next message.
   - The input is formatted with one heading per line, each line starting with a line number enclosed in brackets (e.g., "[123]").
   - The headings are indented to reflect the hierarchy of sections.
(2) Label each heading according to the categories above.
   - Use the glossary below as examples of terms relevant to each category.
   - If multiple categories apply to a section, report all of them in your output.
(3) Report labels for **all** headings in the output as a JSON-formatted string.
   - Format the output as a JSON string containing a list of tuples, with each tuple corresponding to a heading.
   - Each tuple must include the corresponding line number for the heading and its assigned label(s).
   - Print **only** the JSON-formatted string in your output without adding any extra information.

Glossary:

The glossary below includes phrases relevant to each category. This glossary is **not** comprehensive; it is crucial that you also identify relevant phrases not listed below.

- **types**: "Information we collect", "Types of data collected", "Categories of personal data".
- **methods**: "How we collect information", "Data collection methods", "Sources of data we collect".
- **purposes**: "Why do we collect your data", "How we use the information we collect", "Purpose of data collection".
- …

Example: …

(a) Section headings.

Task: Assume the role of a data privacy expert tasked with analyzing website privacy policies. Meticulously extract and catalog specific data types that are mentioned as being collected. Carefully follow the instructions below, using the provided example as a guide.

Instructions:

(1) Carefully and thoroughly read the privacy policy text provided in the next message.
   - The input is formatted with each line starting with a line number enclosed in brackets (e.g., "[123]").
(2) Identify **all** explicit mentions of specific data types or categories that are potentially collected (see the glossary for examples).
   - Identify all mentions regardless of how many times they are repeated throughout the text.
   - Focus on identifying the collected data types and **not** how they are collected and/or used.
   - Ignore mentions in hypothetical or negated contexts, e.g., "we do not collect ...".
   - Separate lists into individual items (e.g., "contact and location information" should be broken down into "contact information" and "location information").
   - Pinpoint the **exact** word(s) used in the text to describe each data type, even if those words are not continuous.
(3) Report the identified data types in the output as a JSON-formatted string.
   - Format the output as a JSON string containing a list of tuples, with each tuple corresponding to an identified data type.
   - Each tuple must include the line number where the data type is mentioned, and the exact word(s) used to describe it in the text (which may be discontinuous).
   - Print **only** the JSON-formatted string in your output without adding any extra information.

Glossary:

The glossary below includes some examples of data types. This glossary is **not** comprehensive; it is crucial that you also identify terms not listed below.

- **Personal Identifier**: "name", "date of birth", "social security number", "driver's license", "passport", "birth certificate", "government-issued identifier", "unique personal identifier"
- **Contact Information**: "email address", "phone number", "postal address"
- **Demographic Information**: "age", "gender", "ethnicity", "marital status", "household data"
- …

Example: …

(b) Collected data types

Figure 2: Example chatbot prompts for labeling section headings extracted from a privacy policy (left) and identifying data types that are mentioned as beling collected (right).

D Full data tables

Table 4: Summary of AI-generated annotations over all categories of collected data types (see Table 1).

	Meta-category	Category	Descriptions
Types (108,748)	Physical profile (36,158)	Contact info (10,582)	email address (27.3%), postal address (25.6%), phone number (25.1%)
		Personal identifier (9,534)	name (31.0%), unique personal identifier (11.7%), social security number (8.6%)
		Professional info (7,779)	employment history (16.3%), employer details (10.8%), job title (10.5%)
		Demographic info (6,203)	gender (14.1%), age (10.6%), demographic info (9.9%)
		Educational info (1,647)	educational info (30.7%), schools attended (6.4%), degrees earned (5.5%)
		Vehicle info (413)	vehicle info (14.3%), vin (10.2%), vehicle registration (5.6%)
	Digital profile (19,182)	Device info (8,659)	browser type (22.4%), operating system (15.6%), device identifier (12.9%)
		Online identifier (4,283)	ip address (65.5%), online identifier (9.1%), domain name (3.9%)
		Account info (3,403)	username (30.1%), password (19.1%), account info (9.0%)
		Network connectivity (1,191)	isp (21.6%), internet connection (17.3%), network traffic (8.0%)
		Social media data (1,040)	social media handle (23.4%), profile picture (19.1%), social media data (9.4%)
		External data (606)	third-party data (24.8%), data from partners (17.2%), inferences (5.6%)
	Bio/health profile (4,751)	Medical info (2,929)	medical info (14.7%), medical conditions (10.1%), disability status (4.3%)
		Biometric data (1,187)	biometric data (25.0%), facial data (12.6%), fingerprint (10.9%)
		Physical characteristic (427)	physical characteristics (46.6%), weight (7.3%), height (6.3%)
		Fitness & health (208)	physical activity info (25.0%), sleep patterns (17.3%), health metrics (3.8%)
	Financial/legal profile (8,864)	Financial info (4,955)	payment card info (25.6%), financial info (15.3%), bank account info (14.7%)
		Legal info (1,729)	signature (21.2%), background checks (9.8%), criminal records (7.2%)
		Financial capability (1,399)	income (17.6%), credit history (13.9%), credit score (7.6%)
		Insurance info (781)	health insurance (29.2%), insurance policy number (19.5%), insurance info (9.7%)
	Physical behavior (4,375)	Precise location (2,389)	gps location (54.8%), precise location (13.0%), device location (4.1%)
		Approximate location (1,620)	country (18.7%), zip code (18.0%), approximate location (17.6%)
		Travel data (276)	movement patterns (26.1%), travel history (10.9%), travel data (2.2%)
		Physical interaction (90)	in-store interactions (43.3%), event participation (4.4%), interactions (4.4%)
	Digital behavior (26,975)	Internet usage (7,847)	browsing history (14.5%), search history (8.3%), click behavior (7.7%)
		Tracking data (3,486)	cookies (43.4%), web beacons (19.0%), online tracking technologies (6.8%)
		Product/service usage (3,076)	user engagement metrics (20.6%), website usage (9.7%), app usage (9.1%)
		Transaction info (2,721)	purchase history (28.6%), transaction info (9.5%), commercial info (5.5%)
		Preferences (2,624)	language preferences (20.3%), preferences (16.5%), product preferences (7.0%)
		Content generation (2,410)	uploaded media (31.7%), comments & posts (9.1%), audio recordings (4.5%)
		Communication data (1,831)	email records (23.4%), call records (15.3%), communication data (9.0%)
		Feedback data (1,259)	survey responses (26.1%), cust. service interactions (13.9%), feedback data (9.9%)
		Content consumption (1,130)	accessed content (62.0%), downloaded content (6.2%), access logs (5.3%)
		Diagnostic data (591)	error reports (13.4%), crash reports (10.7%), diagnostic data (9.1%)

Table 5: Breakdown of collected data types over all categories (see Table 2).

Meta-category	Category	Overall statistics		Sector statistics (sorted by coverage)											
		Coverage	Mean/SD	Highest			2nd highest			3rd highest			Lowest		
Physical profile	Contact info	86.4%	3.6±1.4	HC	91.0%	3.5±1.3	TC	90.8%	3.7±1.0	CD	90.4%	3.8±1.2	FS	77.4%	3.4±1.6
	Personal identifier	89.5%	3.4±2.6	TC	93.9%	3.3±2.2	CD	91.8%	3.8±2.6	CS	91.3%	3.5±2.4	EN	77.8%	2.6±2.1
	Professional info	59.0%	4.5±5.0	IT	68.7%	5.1±5.6	HC	65.6%	4.8±4.9	TC	65.3%	3.9±4.7	UT	44.4%	3.0±2.9
	Demographic info	49.9%	4.7±4.2	TC	67.3%	4.2±3.8	CD	65.3%	4.7±4.0	CS	62.1%	4.9±4.0	MT	29.8%	3.9±4.1
	Educational info	27.9%	2.2±2.3	HC	34.6%	1.7±1.3	FS	31.4%	2.5±2.3	CS	28.2%	2.0±2.2	MT	15.8%	2.4±2.8
	Vehicle info	5.0%	3.0±8.2	CD	11.3%	5.6±15.5	RE	9.7%	1.4±0.5	IN	8.0%	2.3±2.1	HC	0.4%	2.0±1.4
Digital profile	Device info	74.4%	4.0±2.9	TC	88.8%	4.6±2.9	CD	86.3%	4.5±3.5	IT	83.0%	4.3±3.2	FS	58.3%	4.0±2.5
	Online identifier	80.9%	1.7±0.9	TC	88.8%	1.9±1.5	CD	88.3%	1.9±1.1	UT	87.0%	1.3±0.8	FS	65.7%	1.7±0.9
	Account info	50.0%	2.4±1.6	CD	64.6%	2.5±1.7	TC	62.2%	2.3±1.5	IT	60.4%	2.4±1.6	EN	30.3%	2.2±1.6
	Network connectivity	29.5%	1.5±1.0	CD	45.0%	1.5±1.1	TC	44.9%	2.3±1.6	IT	34.7%	1.6±1.1	EN	14.1%	1.4±0.6
	Social media data	23.3%	1.6±1.2	CD	39.5%	1.7±1.4	TC	36.7%	2.3±1.5	CS	34.0%	1.8±1.4	MT	9.6%	1.2±0.4
	External data	12.4%	1.7±1.4	TC	23.5%	1.7±1.2	UT	18.5%	1.4±1.0	CS	17.5%	1.3±0.6	EN	5.1%	1.0±0.0
Bio/health profile	Medical info	28.3%	3.7±3.5	HC	50.1%	4.7±4.4	CS	31.1%	3.6±2.7	FS	28.0%	4.0±3.8	EN	11.1%	1.9±1.6
	Biometric data	16.4%	2.6±3.0	FS	20.2%	3.6±3.8	HC	19.1%	2.4±2.9	CD	18.9%	2.3±2.2	EN	3.0%	2.7±2.9
	Physical characteristic	11.2%	1.5±1.1	CS	16.5%	1.6±1.1	FS	16.1%	1.4±0.9	CD	14.4%	1.8±1.6	EN	4.0%	1.0±0.0
	Fitness & health	3.5%	2.2±2.5	TC	7.1%	1.7±1.5	CD	5.2%	3.5±4.0	HC	4.7%	2.0±1.9	IT	1.5%	1.4±0.9
Financial/ legal profile	Financial info	53.9%	3.2±2.3	CD	73.5%	3.3±2.1	UT	64.8%	2.6±1.9	FS	63.9%	3.5±2.9	EN	27.3%	2.7±1.5
	Legal info	28.7%	2.3±2.1	FS	35.9%	2.7±2.6	CD	33.0%	2.0±1.7	RE	32.3%	2.5±1.7	MT	16.7%	1.6±1.1
	Financial capability	21.5%	2.5±2.1	FS	51.6%	3.1±2.2	RE	22.6%	2.6±1.6	CD	19.2%	2.6±2.3	CS	8.7%	1.2±0.4
	Insurance info	14.8%	2.0±1.7	FS	24.2%	2.9±2.6	HC	22.2%	1.6±1.2	CD	13.4%	1.5±0.6	MT	6.1%	2.0±0.0
Physical behavior	Precise location	50.9%	1.5±0.9	TC	71.4%	1.6±1.1	CD	68.4%	1.7±1.1	CS	59.2%	1.6±0.9	EN	25.3%	1.4±0.6
	Approximate location	33.3%	1.8±1.2	TC	54.1%	2.0±1.5	IT	44.9%	1.9±1.2	CD	43.0%	1.9±1.2	UT	16.7%	1.1±0.3
	Travel data	6.6%	1.6±1.9	IN	10.4%	2.0±3.0	CD	9.6%	2.0±1.9	TC	9.2%	2.3±2.5	UT	1.9%	2.0±0.0
	Physical interaction	2.8%	1.2±0.5	CD	6.5%	1.0±0.0	RE	4.0%	1.8±0.8	IN	3.6%	1.0±0.0	FS	1.6%	1.0±0.0
Digital behavior	Internet usage	72.8%	3.8±2.8	TC	84.7%	4.1±2.9	CD	83.2%	4.4±3.1	CS	80.6%	4.0±2.3	EN	48.5%	3.1±2.5
	Tracking data	46.7%	2.3±1.6	CD	55.0%	2.3±1.6	IT	54.2%	2.2±1.6	TC	51.0%	2.7±2.0	FS	37.7%	2.4±1.6
	Product/service usage	50.8%	2.1±1.8	TC	72.4%	2.4±1.8	CD	61.9%	2.5±2.6	CS	60.2%	1.9±1.2	EN	32.3%	2.2±1.7
	Transaction info	43.9%	2.2±1.5	CD	63.9%	2.7±2.1	FS	60.1%	2.1±1.6	CS	58.3%	2.6±1.5	EN	21.2%	2.0±1.2
	Preferences	49.1%	2.0±1.3	CD	65.6%	2.4±1.7	CS	64.1%	2.1±1.4	TC	54.1%	2.2±1.6	UT	29.6%	2.0±0.8
	Content generation	32.8%	2.3±1.9	CD	49.5%	2.5±1.8	TC	41.8%	2.3±1.4	CS	41.7%	2.7±2.2	UT	13.0%	1.3±0.5
	Communication data	33.8%	1.9±1.4	TC	48.0%	2.0±1.4	CD	42.6%	1.9±1.4	IT	39.0%	2.1±1.6	UT	11.1%	1.8±1.0
	Feedback data	25.3%	1.8±1.2	CD	37.1%	2.1±1.6	CS	34.0%	1.6±0.9	IT	31.0%	1.9±1.2	EN	12.1%	1.9±1.6
	Content consumption	26.7%	1.3±0.8	TC	46.9%	1.9±1.2	IT	34.7%	1.5±1.2	CS	33.0%	1.1±0.2	UT	11.1%	1.0±0.0
	Diagnostic data	14.3%	1.6±1.3	TC	26.5%	1.5±0.9	IT	22.0%	2.0±1.7	IN	17.1%	1.6±1.7	EN	4.0%	1.0±0.0

Table 6: Examples of validated AI-generated annotations and the validation context. Note that there can be multiple annotations from the same context, e.g., when multiple collected data elements are reported in the same sentence.

	Category	Descriptor	Text	Context
Types	Biometric data	retina scan	imagery of the iris or retina	Biometric Information, such as voice prints, imagery of the iris or retina, face geometry, and palm prints or fingerprints
	Demographic info	citizenship	citizenships held	Passport details, place of birth, citizenships held (past and present), and residency status
	Device info	browser type	type of browser software	X logs your current Internet address (this is usually a temporary address assigned by your Internet service provider when you log in), the type of operating system you are using, and the type of browser software used.
	Financial capability	student loan information	student loan financial information	Information regarding your education history, including degrees earned and student loan financial information.
	Precise Location	gps location	latitude and longitude coordinates	X collects latitude and longitude coordinates from the device as part of the timekeeping process when geolocation services are enabled
	Product/service usage	website usage	use of our website	For example, from observing your actions as a candidate, from records of your use of our website, network, or other technology systems.
Purposes	Basic functioning	contract fulfillment	For the performance of a contract or to conduct business with you	For the performance of a contract or to conduct business with you (e.g., consulting; speaker agreement).
	Data sharing	data sharing with affiliates	provide personal information to our affiliated businesses	To the extent permitted by applicable law, we may provide personal information to our affiliated businesses or to our business partners, who may use it to send you marketing and other communications.
Handling	Data retention	Stated	retain your personal information for the period you are actively using our services plus six (6) years	We retain your personal information for the period you are actively using our services plus six (6) years.
	Data protection	Generic	commercially reasonable administrative, technical, and organizational safeguards	We strive to protect the information you provide to us when you use our X Services through commercially reasonable administrative, technical, and organizational safeguards.
	Data protection	Secure transfer	Secure Socket Layer (SSL) encryption technology for payment transactions	Steps we have taken to enhance network and information security include industry standard infrastructure security, the implementation of Secure Socket Layer (SSL) encryption technology for payment transactions, digital certificates, and ...
Rights	User choices	Privacy settings	change your preferences as well as update your Personal Information through your account settings	If you have a registered account, you may be able to change your preferences as well as update your Personal Information through your account settings.
	User choices	opt-out via link	click the Opt-Out of Sale/Sharing Request tab on this page	To submit a request to opt out of the sale or sharing of your personal information, please click the Opt-Out of Sale/Sharing Request tab on this page.
	User access	Edit	see and/or update certain of your personal information	We offer various self-help tools that will allow you to see and/or update certain of your personal information in our records.