Asking for Privacy: Contrasting Consumer Questions with Questions in Privacy Policies

Tianyang Zhao
Pennsylvania State University
University Park, PA, USA
tzz4@psu.edu

Younes Karimi
Pennsylvania State University
University Park, PA, USA
younes@psu.edu

Thomas Norton
Fordham University School of Law
New York, NY, USA
tnorton1@law.fordham.edu

Norman Sadeh
Carnegie Mellon University
Pittsburgh, PA, USA
sadeh@cs.cmu.edu

Kami Vaniea
University of Waterloo
Waterloo, ON, Canada
kami.vaniea@uwaterloo.ca

Shomir Wilson
Pennsylvania State University
University Park, PA, USA
shomir@psu.edu

Abstract

Privacy policies are text documents intended to inform consumers about the data practices of apps and websites, but they can be challenging to read. Some organizations try to address the obstacles by organizing their privacy policies as question-answer pairs. We use a combination of automated and manual analysis methods to compare two corpora: PrivaSeerQA, a corpus we create and release consisting of 1.2 million question-answer pairs extracted from 345,073 privacy policies, and PrivacyQA, a previously released corpus of 1,750 questions people ask about app privacy. We find a mixture of thematic overlaps (e.g., data types and data uses) and divergences (e.g., levels of granularity and the issue of location tracking). We also find that privacy policies have a greater focus on the visibility of data, while users have a greater focus on their data being sold. The contrasts suggest that QA-structured privacy policies still do not fully meet consumers' needs, although the structure clarifies what specific topics of user interest are missing.

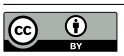
CCS Concepts

• **Computing methodologies** → **Language resources**; • **Security and privacy** → **Usability in security and privacy**.

Keywords

privacy policy, privacy, legal, natural language processing, corpus, large language model

ACM Reference Format:
Tianyang Zhao, Younes Karimi, Thomas Norton, Norman Sadeh, Kami Vaniea, and Shomir Wilson. 2026. Asking for Privacy: Contrasting Consumer Questions with Questions in Privacy Policies. In *Proceedings of the 2026 ACM Symposium on Document Engineering (DocEng '26), August 25–28, 2026, Fribourg, Switzerland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3820755.3832806



This work is licensed under a Creative Commons Attribution 4.0 International License.
DocEng '26, Fribourg, Switzerland
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2786-3/2026/08
https://doi.org/10.1145/3820755.3832806

1 Introduction

Privacy policies are intended to provide transparency about organizations' data practices, and consumers are increasingly aware of the potential consequences of personal data collection and usage. However, privacy policies represent a *lemon market* [19] where organizations have detailed knowledge about how they collect and use data but consumers have only the limited information provided in privacy policies. Additionally, many privacy policies are long, involving ambiguous language and technical vocabulary [10, 14]. Consumers often feel overwhelmed by privacy policies to the extent they do not open them or read them [12]. Further, consumers who do open them typically spend fewer than two minutes on reading [3], less time than the estimated 8-12 minutes this would require [10]. This situation creates a paradox in the general public's attitude toward privacy: they wish to learn about organizations' data practices, but they often agree to policies they never read [12].

In response to the obstacles of understanding privacy policies, organizations often include text formatting that signposts topics, potentially improving readability. Such a structure proactively labels different sections, but it also represents an unclear intersection between the information the organization wants to convey, legal requirements, and assumptions surrounding consumers' interests. In this paper we take advantage of the structure organizations give privacy policies to study at scale the types of information they intend to convey to people. Specifically, we investigate the research question: *From a qualitative perspective, how do questions in question-answer (QA) pairs extracted from privacy policies differ from privacy questions asked by consumers?* We start by creating and releasing PrivaSeerQA[1], a new corpus of 1,293,714 English language question-answer pairs (QA pairs) extracted from the the PrivaSeer Corpus [17]. This is the largest privacy-oriented question answering corpus to date. We conduct a thematic analyses using large language model-supported methods (specifically Llama 3.3 [5]) and manual methods to compare PrivaSeerQA with PrivacyQA, a previously released corpus of 1,750 questions users ask about app privacy [13], showing their key differences. We further intend for the PrivaSeerQA corpus to enable the construction of human language technologies for consumers, privacy researchers, and privacy scholars to improve digital privacy.

[1] https://git.psu.edu/hlt-lab/PrivaseerQA-Corpus

2 Related Work

Privacy policies are time-consuming to read and difficult for end-users to comprehend. McDonald et al. [11] investigated the obstacles to understanding privacy policies using quantitative and qualitative studies. They found that privacy policies had Flesch Reading Ease scores from 32 to 46, suggesting a need for undergraduate-level reading skills to understand these documents, thus indicating low readability. Amos et al. [2] found that privacy policies became much longer from 2009 to 2019, with the median word count doubling. They also found many privacy policies included links to other privacy policies, increasing the reading time for users.

Question answering on privacy policies has thus gained attention as a way to ease the understanding of their contents. Harkous et al. [7] created *Polisis*, chatbots designed to answer questions about privacy practices by automating information extraction and text generation about them. PrivacyQA [13] and PolicyQA [1] were later introduced as the first dedicated privacy policy question answering corpora. PrivacyQA focused on mobile privacy policies, by asking crowdworkers to write questions about app behavior and legal experts to find answers in corresponding apps' privacy policies. PolicyQA focused on web privacy policies, by automatically generating question-answer pairs from annotations in the OPP-115 Corpus [20]. PrivacyQA and PolicyQA were widely used in later studies. For example, Shankar et al. [16] used both in a benchmark for evaluating language models' understanding of privacy policies.

3 PrivaSeerQA Corpus Creation

We performed an informal manual examination of a small number of privacy policies structured as QA lists and observed:

(1) Question marks usually indicate the end of a question.
(2) Questions shorter than three words (e.g., *what are cookies*) are exceedingly rare.
(3) The answer to a question typically follows the question, immediately and without intervening words.

To extract QA pairs from a privacy policy, we started by segmenting it into sentences, and then we automated observations (1) and (2) to identify questions. Observation (3) was used to identify answers, with each answer being a sequence of one or more consecutive sentences ending before the next question. We performed this procedure on the entirety of the PrivaSeer Corpus, noting that some privacy policies contained zero questions. Accordingly, PrivaSeerQA's 1.2 million questions represent the contents of 345,073 privacy policies in the PrivaSeer Corpus. Due to its large size, PrivaSeerQA was divided into 50 *chapters*, with each chapter containing the questions found in a subset of privacy policies from PrivaSeer. Across chapters, the ordering of privacy policies is the same as in PrivaSeer, which placed them in random order.

4 Thematic Analysis of Questions

Subsets of 200 unique questions from both corpora were created by randomly sampling from them after de-duplication (internally to each corpus) of their contents. Analyses of these sets were performed separately, so the findings would best represent the respective corpora rather than requiring the two corpora to conform to the same observations. Using qualitative study methods, the lead researcher conducted an affinity analysis for each subset of questions while also taking memos. A second researcher then reviewed and refined the analysis with the lead researcher in a multi-hour meeting where memos were also read. When sharing examples from the corpora, we make minor adjustments for readability.

Table 1: Themes in PrivacyQA questions asked by users, with high-level themes in grey and sub-themes in white.

Question	Count (%)
Who has access to data	65 (33.9%)
Selling Data	19
Third party use/sharing of data	18
Who can see my data or conclusions drawn about me?	17
Data access by peers	11
Data collection and use	64 (33.3%)
Access and saving of data obtained from phone	27
What data is collected?	11
How is information used?	10
Is location tracked?	10
Is browsing activity tracked?	6
User control, consent, and awareness	20 (10.4%)
User controls and consent	10
Making user aware of data collection	7
User requirements around data accuracy	2
Right to be let alone	1
Security	27 (14.1%)
Security of the app and information	20
Payment information	3
Location of data storage	2
Kids' safety	2
Retention	9 (4.7%)
Information collected about other people	2 (1%)

Table 1 shows the hierarchy of themes found in PrivacyQA questions. Of the 200 questions, 8 were not about privacy. Percentages in the table are out of 192 and are to help the reader understand the relative frequency of question types. Table 2 shows the hierarchy of themes found in privacy policy questions from PrivaSeerQA. Of the 200 questions sampled, 28 were not about privacy (e.g., *"FREE estimate on your Shower Glass Panel project?"*) or contained insufficient context to interpret (e.g., *"How do they work?"*).

We use the Llama 3.3 [5] Large Language Model (LLM) with one-shot learning to apply the themes from the qualitative analysis above to two chapters of PrivaSeerQA (1 and 36), which contained in aggregate 52,041 questions. We provide 23 one-shot examples, one for each theme, and asked the model to assign the most fitting theme to each question from PrivaSeerQA or declare the question fits none of the themes. This allows us to check that the 200 randomly selected questions are indeed representative.

Additionally, we examined the reliability of this automatic thematic coding approach by using it on the remaining 177 questions in the set of 200 sampled questions (i.e., 200 minus the 23 provided to the model as one-shot examples). Cohen's kappa between Llama 3.3's responses and manual analysis was 0.7586, suggesting a level of agreement that supports labeling of the rest of the corpus.

Table 2 shows the results of automated labeling. The relative proportions for each theme are roughly similar across the manual

Table 2: Comparison of PrivaSeerQA theme prevalence in Llama3.3 labeling and manual thematic analysis. Counts are of questions labeled with that theme. Higher-level themes are shown in grey with sub-themes in white below them.

Question	Ch. 1	Ch. 36	Manual
Data collection and use	11668 (47.4%)	11780 (47.2%)	76 (44.2%)
How is information used?	3069	3041	24
What information is collected?	3059	3056	15
What information is collected and how is it used?	502	502	3
Why is the data needed?	917	1009	10
Cookies	2563	2595	9
How is data collected?	1294	1311	6
Children	264	266	1
Consent and user control	3800 (15.4%)	3782 (15.2%)	32 (18.6%)
Removing/expressing consent	1725	1688	12
Right to access and correct	1356	1399	11
Consent and legal basis	719	695	9
Who can see data and who is it shared with	3351 (13.6%)	3507 (14.1%)	23 (13.4%)
Who is data shared with?	1169	1258	15
Sharing with third parties	1709	1780	5
Sharing with advertisers	473	469	3
Data security	2736 (11.1%)	2808 (11.3%)	19 (11.0%)
How is data protected?	2279	2345	12
Location of data storage or processing	323	331	3
Payment information	50	42	2
Third-party security	84	90	2
Legal mechanics	1328 (5.4%)	1377 (5.5%)	15 (8.7%)
Changes to privacy policy	395	411	7
Legal definitions	523	525	4
Scope of privacy policy	410	441	4
Contact the company	978 (4.0%)	935 (3.7%)	9 9 (5.2%)
Questions or concerns	550	508	5
How to contact	428	427	4
Retention	757 (3.1%)	745 (3.0%)	8 (4.7%)

and automated sets, suggesting our sample questions represent the dataset well. 4.6% of questions were labeled by the LLM as fitting no themes, and manual review confirmed this.

5 Discussion

Data collection and use is the most popular topic in privacy policy questions (~47%), and it appears frequently in user questions (~33%). Data access is another top interest of users (i.e., ~34% of PrivacyQA), but it only appears in ~14% of privacy policies. On the other hand, *consent and user control* is the second most popular topic in privacy policies (~15%), but only a modest number of user questions (~10%) concern it. Below we discuss findings on selected themes.

5.1 Selling Versus Sharing Data

Both our own analysis of PrivacyQA and the analysis by Zhang and Sadeh [21] found that many people ask questions about *selling* data. Privacy policies, by comparison, tend to use terms like "share" or "see" in their questions for the same activities. This framing could reflect attempts by companies to soften the concept of data selling (i.e., in the context of data brokers). However, given the imprecision of these terms and consumers' limited legal knowledge, this alternative terminology is likely to be confusing for consumers.

5.2 Terms and Precision

User questions often were precise about data types and types of usage, e.g.,: *"Does it save my email contacts and sell it to 3rd parties?"* Zhang and Sadeh [21] hypothesized that definitions of words like "collect", "store", and "save" are unclear to users, and the definitions in the Google Play store likely differ from user expectations. For example, information that is accessed on-device but never leaves it is not considered to be "collected", which may not be intuitive to users. Similarly, we observed users' efforts to clarify terms in their questions, like: *"Does any human ever see the images that I choose to send?"* Such clarifications are infrequent in the questions found in privacy policies, and although they may occur in answers, a user browsing the QA structure could miss that additional information.

5.3 Perspective and Terminology

The difference in perspective between users and organizations is also apparent between PrivacyQA and PrivaSeerQA, respectively. PrivacyQA questions show that users sometimes frame questions conditionally to clarify their concerns. For example: *"Is there money exchanging hands if my info is shared?"* Similarly, the question *"Does any human ever see the images that I choose to send?"* indicates concern specifically when pictures are sent. We also encounter questions that show interest in situations when a human may perceive data differently than an automated process: *"If I send a message that is considered dirty, will the controllers of the app see it?"* In contrast, PrivaSeerQA questions appear to lack these nuances. We also see many cases when users are vague about the meanings of terms, especially about entities accessing their data. For example, the following quote refers to "people" but does not distinguish between other users of the app or people associated with the organization: *"Can people see what celebs I have looked up?"*

Privacy policy questions appear to draw from a narrower set of language than users' questions, reflecting the precision of legal text in the privacy domain. Even in a sample of only 200 questions, we see many near-duplicate questions that use the same key terms like "personal information". For privacy professionals, key terms in privacy policies are often assumed to have specific recognized meanings [8, 9]. However, prior work has also found that interpretations of privacy policies, particularly around data sharing practices, can vary between experts and end users [15, 18]. Tang et al. tested 22 privacy policy terms to understand if users correctly interpret them [18], and we find four of their terms in our set of 200 manually examined PrivaSeerQA questions: "tracking" (understood by 51% of Tang et al.'s participants), "Do Not Track" (36%), "personal information" (86%), and "personally identifiable information" (69%). While confusion likely remains, it is encouraging to see the general avoidance of confusing terms in privacy policy questions.

5.4 Privacy from Peers

Regulations and laws like the General Data Protection Regulation [6] often focus on how organizations use the data they collect from their users. However, users are also concerned about data access and usage by their peers (i.e., fellow users), including scenarios

when other users within the same household space pose a threat to their personal safety [4]. Privacy policies rarely mention peers, and instead tend to focus on business-to-business (B2B) sharing or visibility to employees. For example: *"What information is collected when a company uses Adobe Experience Cloud?"* This is a notable gap in the status quo of privacy policy contents.

5.5 Data Collection Mechanisms

"How" was the most common word to begin questions in PrivaSeerQA, but questions about how data were collected were notably rare. Privacy policy questions barely addressed data collection mechanisms, even though they often addressed data types and data usage. In contrast, users represented in PrivacyQA wanted to know details like whether data would be collected from the phone on which an app was installed, or if data could be collected from their friends.

6 Limitations

The question identification and answer identification methods in the PrivaSeerQA creation pipeline contain heuristic rules that are grounded in English language norms, but exceptions exist. Manual observations found occasional errors but not pervasive ones, and the scalability of our automated methods was necessary for this analysis of privacy policies. Additionally, PrivacyQA, as a set of human-generated questions chiefly provoked by the descriptions of Google Play Store apps, could be more focused on app-related issues. While our analysis confirms that the two corpora merit comparison, it is possible that a corpus of human-generated questions about websites could contain different themes.

7 Conclusion

We presented PrivaSeerQA, a new QA corpus extracted from privacy policies. Its breadth and structure permit new observations on the contents of privacy policies, and we release the corpus for the benefit of other privacy researchers. Our qualitative analysis of a subset of PrivaSeerQA and PrivacyQA questions revealed similarities and differences between the questions organizations proactively answer and the questions users wish to ask. LLM-based coding of additional PrivaSeerQA questions compared favorably to manual results, supporting the subset analysis. The differences between QA-structured privacy policies and users' questions show that QA-structured privacy policies still do not meet users' needs, with some signs of intentional obfuscation (e.g., *sell* versus *share*). However, the contrast points to specific topics that privacy policies can improve upon for greater alignment with users' interests.

8 Acknowledgments

This manuscript is based upon work supported by the National Science Foundation under grant Nos. 1914444, 1914446, 1914486, 2105736, and 2237574 and NSERC grant No. 2024-06737. This work benefited from Dagstuhl Seminar 25021 "Grand Challenges for Research on Privacy Documents".

References

[1] Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A Reading Comprehension Dataset for Privacy Policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 743–749.

[2] Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, 2165–2176.

[3] Yannis Bakos, Florencia Marotta-Wurgler, and David R. Trossen. 2014. Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts. *The Journal of Legal Studies* 43, 1 (2014), 1–35.

[4] Alaa Daffalla, Arkaprabha Bhattacharya, Jacob Wilder, Rahul Chatterjee, Nicola Dell, Rosanna Bellini, and Thomas Ristenpart. 2025. A framework for abusability analysis: the case of passkeys in interpersonal threat models. In *Proceedings of the 34th USENIX Conference on Security Symposium* (Seattle, WA, USA) *(SEC '25)*. USENIX Association, USA, Article 401, 20 pages.

[5] Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[6] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council.

[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 *Proceedings of the 27th USENIX Conference on Security Symposium (SEC'18)*. USENIX Association, USA, 531–548.

[8] Yang Liu, Alan Medlar, and Dorota Głowacka. 2022. Lexical Ambiguity Detection in Professional Discourse. *Information Processing & Management* 59, 5 (Sept. 2022), 103000.

[9] Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 6929–6947.

[10] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society* 4 (2008), 543.

[11] Aleecia M. McDonald, Robert W. Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. 2009. A comparative study of online privacy policies and formats. In *Proceedings of the 5th Symposium on Usable Privacy and Security (SOUPS '09)*. Association for Computing Machinery, New York, NY, USA, Article 46, 1 pages.

[12] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147.

[13] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4947–4958.

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

[15] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, Rohan Ramanath, et al. 2015. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ* 30 (2015), 39.

[16] Atreya Shankar, Andreas Waldis, Christof Bless, Maria Andueza Rodriguez, and Luca Mazzola. 2023. PrivacyGLUE: A Benchmark Dataset for General Language Understanding in Privacy Policies. *Applied Sciences* 13, 6 (2023).

[17] Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6829–6839.

[18] Jenny Tang, Hannah Shoemaker, Ada Lerner, and Eleanor Birrell. 2021. Defining Privacy: How Users Interpret Technical Terms in Privacy Policies. *Proceedings on Privacy Enhancing Technologies* (2021).

[19] Tony Vila, Rachel Greenstadt, and David Molnar. 2003. Why we can't be bothered to read privacy policies models of privacy economics as a lemons market. In *Proceedings of the 5th International Conference on Electronic Commerce (ICEC '03)*. Association for Computing Machinery, New York, NY, USA, 403–407.

[20] 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, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1330–1340.

[21] Shikun Zhang and Norman Sadeh. 2023. Do privacy labels answer users' privacy questions?. In *Workshop on Usable Security and Privacy*.