

Incorporating Taxonomic Reasoning and Regulatory Knowledge into Automated Privacy Question Answering

Abhilasha Ravichander^{*1}, Ian Yang^{*2,3}  Rex Chen³, Shomir Wilson⁴ ,
Thomas Norton⁵, and Norman Sadeh³ 

¹ University of Washington
aravicha@cs.washington.edu

² Georgia Institute of Technology
iyang30@gatech.edu

³ Carnegie Mellon University
rexc@cmu.edu, sadeh@cs.cmu.edu

⁴ Penn State University
shomir@psu.edu

⁵ Fordham University
tnorton1@law.fordham.edu

Abstract. Privacy policies are often lengthy and complex legal documents, and are difficult for many people to read and comprehend. Recent research efforts have explored automated assistants that process the language in policies and answer people’s privacy questions. This study documents the importance of two different types of reasoning necessary to generate accurate answers to people’s privacy questions. The first is the need to support taxonomic reasoning about related terms commonly found in privacy policies. The second is the need to reason about regulatory disclosure requirements, given the prevalence of silence in privacy policy texts. Specifically, we report on a study involving the collection of 749 sets of expert annotations to answer privacy questions in the context of 210 different policy/question pairs. The study highlights the importance of taxonomic reasoning and of reasoning about regulatory disclosure requirements when it comes to accurately answering everyday privacy questions. Next we explore to what extent current generative AI tools are able to reliably handle this type of reasoning. Our results suggest that in their current form and in the absence of additional help, current models cannot reliably support the type of reasoning about regulatory disclosure requirements necessary to accurately answer privacy questions. We proceed to introduce and evaluate different approaches to improving their performance. Through this work, we aim to provide a richer understanding of the capabilities automated systems need to have to provide accurate answers to everyday privacy questions and, in the process, outline paths for adapting AI models for this purpose.

Keywords: privacy · question answering · generative AI · regulation

^{*} The first two authors contributed equally to this work. Please refer questions about this research to the last author, Prof. Norman Sadeh - sadeh@cs.cmu.edu

1 Introduction

Privacy policies are legal documents describing how user data is collected, used, managed and shared. In practice, these documents are difficult for users to read and understand. In response, researchers have advocated the adoption of privacy labels as succinct summaries of data practices [6, 7]. Yet, research shows that labels produced so far remain too complex to be usable and fail to address a significant percentage of typical privacy questions [17]. There has been research interest in automatically answer people’s privacy questions, rather than limiting users to a fixed set of questions, as privacy labels do [5, 12].

In this work, we explore the capabilities required to support accurate automated privacy question-answering. Specifically, we collected fine-grained semantic annotations of privacy policies, using annotators with a legal background to identify privacy policy statements relevant to answering typical mobile app privacy questions. Collectively our expert annotators produced collections of annotations for a total of 210 privacy policy/question pairs, with each privacy policy/question pair annotated by at least 3 different annotators for a total of 749 collections of annotations. Detailed analysis of these annotations sheds light on the importance of taxonomic reasoning and/or reasoning about regulatory disclosure requirements to generate accurate answers to people’s everyday privacy questions. In a second part of this research, the annotations are used as ground truth to evaluate the ability of generative AI tools to perform the type of taxonomic and regulatory reasoning required to accurately answer many typical privacy questions. While in their default configuration these tools struggle to correctly answer many questions, we demonstrate prompting strategies that augment their default knowledge, and show that this can lead to a relative improvement of 8.75% in GPT-4 performance.

This work documents the importance of taxonomic reasoning and reasoning about regulatory disclosure requirements in answering common privacy questions. Privacy policy texts often use terms that may not exactly match those in a user’s question or those found in applicable regulations. Understanding the relationships between these terms is critical to determining statements relevant to answering a particular question, and often requires domain knowledge (e.g., a policy statement about "GPS data" or the particular "stores a user visits" might be relevant to a question about whether a mobile app collects the user’s "location"). Privacy QA also needs to address *policy silence*, namely common situations where the text of a privacy policy simply does not address a given question (e.g., no statement about whether an app collects or does not collect the user’s location information). The absence of a statement about whether an entity engages in a given data practice (e.g., sharing one’s location with third party advertisers) is to be interpreted differently depending on whether applicable regulation require the disclosure. In contrast, silence when applicable regulations do not require such a disclosure does not have the same implication.

The contributions of this work can be summarized as follows: (1) We show the importance of taxonomic reasoning and reasoning about regulatory disclosure requirements in accurately answering common privacy questions. This includes

reporting on the wide range of phrases commonly used in policy text and the need to explicitly reason about their subsumptive relationships. This further includes reporting on the prevalence of silence in the text of privacy policies and the importance of interpreting silence in the context of applicable regulatory disclosure requirements; (2) We evaluate the ability of generative AI to answer privacy questions and show that they generally seem oblivious to regulatory disclosure requirements; (3) We show how using prompting techniques designed to (a) provide generative AI tools with taxonomic and regulatory knowledge they are missing and (b) encourage them to break down reasoning and explicitly identify silence in policies, it is possible to significantly enhance the performance of state-of-the art generative AI tools.

2 Related Work

Privacy policies have been a growing area of interest in NLP research, given their wide availability, their increasing complexity but also society’s growing concerns about privacy issues and the emergence of increasingly stringent privacy regulations [9, 14, 16, 18]. Given that few people ever have the time to read privacy policies, automated privacy question answering has emerged as a promising approach to empowering people to take advantage of the disclosures found in the text of policies without requiring them to actually read the policies. The PrivacyQA Corpus [12] consists of 1,750 crowdsourced questions about mobile apps’ privacy behaviors, which the corpus creators used to train a QA system. Other early work on privacy question answering has been reported by Harkous et al. in the context of *PriBot* [5].

Oltramari et al. report on their work on PrivOnto, a semantic framework for modeling and reasoning about privacy practices [10]. Bhatia et al. report on their work to automate the extraction of regulated information types using hyponymy relations [3]. Evans et al. discuss the identification of Tregex patterns to automate the extraction of hyponyms from the text of privacy policies [4]. [1] seek to identify contradictions in privacy policy text, and [8] describe the granularity at which statements in smart home privacy policies are described.

In this paper, we study the inferences a question-answering (QA) system must make to accurately answer common privacy questions people have, document the importance of both taxonomic reasoning and reasoning about regulatory data practice disclosure requirements, evaluate the ability of state-of-art generative AI tools to do this type of reasoning and discuss ways of enhancing the performance of these tools.

3 Methodology

This study aims to document the importance of taxonomic inference and of reasoning about regulatory disclosure requirements in answering privacy questions. We describe the annotation process below:

Question Selection Our focus is on automating the generation of answers to well-formed privacy questions over mobile apps. We opted to disregard complexities related to how people phrase their privacy questions (see [12] for a study of these issues). We qualitatively analyzed a corpus of privacy questions [12]. We sampled 365 questions and for purposes of our analysis, disregarded generic questions such as ‘Does this app collect my information?’, as these types of questions are overly vague. We found that common questions focused on a small set of data practices such as the collection or sharing of different types of sensitive data. This includes location information (27.4% of the questions in our sample), camera access (6.57%), credit card information (3%), and contacts list (5.48%). Thus, we constructed seven prototypical mobile app privacy questions: 3 questions about the collection of sensitive data and 4 dealing with the sharing of sensitive data. One question singles out sharing with advertisers, a common privacy concern. The resulting questions are: (Q1): Does this app collect my location information?, (Q2): Does this app access my camera?, (Q3): Does this app collect credit card information?, (Q4): Does this app share my location with others? (Q5): Does this app share my location with advertisers? (Q6): Does this app share my contacts list? (Q7): Does this app share my credit card information?

Mobile App Privacy Policy Selection Policies in the PrivacyQA corpus span 10 categories of apps. After piloting our annotation process, it was determined that collecting annotations for our seven questions for one privacy policy would take about an hour. With a pool of 7 annotators able to commit about 5 hours of their time each week, we estimated that we would be able to annotate about 11 mobile app privacy policies per week. Taking into account delays, we estimated that in a month we would be able to annotate about 30 privacy policies - with 3 sets of annotations for each policy. Accordingly, we sampled six categories (out of the 10): health, travel, news and magazines, entertainment, lifestyle and games. Within each category, we chose three apps with more than 10 million downloads (popular apps) and two apps with fewer than 10 million downloads, on the Google playstore. We deliberately excluded apps with fewer than 100,000 downloads, as we wanted to avoid apps with privacy policies of possibly uneven quality. All policies were collected between August and October 2022.

Modeling Relationships Between Relevant Terms One challenge in answering privacy policy questions arises from the diversity of terms that can possibly be used to refer to related concepts such as related data types and data practices. Specifically, one can formulate the problem of answering a privacy question based on the text of a privacy policy as a process that involves identifying in the text of the policy relevant reference frames related to the data practice discussed in the question (e.g., "Does this app collect my location?", "Does this app share my location with third parties?"). As discussed in the previous section, in this study, we focus on questions related to the collection and sharing of different data types. Relevant reference frames in the text of a privacy policy will be text fragments that refer to the data practice and data type discussed in the question,

whether directly (i.e., using the exact same terms) or indirectly (i.e., by using related terms or phrases). This is defined below.

Definition 1 (Evidence Frame). An evidence frame $e=(i,a,r,m)$, where i is the information type, a is the action performed on that information type such as collection/sharing, r is the information recipient, and m is the modality ($m \in \{performed, not_performed, performed_under_specified_condition, may_be_performed\}$.)

Question	Segment	Relation
Does this app collect my [location information] $_{i_q}$?	This app will collect [GPS information] $_{i_p}$.	$\subset$
Does this app collect my [location information] $_{i_q}$?	This app will collect [location information] $_{i_p}$.	$\equiv$
Does this app collect my [location information] $_{i_q}$?	This app will collect [personal information] $_{i_p}$ such as email addresses, phone numbers etc.	$\supset$
Does this app collect my [health information] $_{i_q}$?	This app will ask for [information about how you get to work, and the distance between home and work.] $_{i_p}$.	$\sqsubseteq$
Does this app collect [information about visits to abortion clinics] $_{i_q}$?	This app uses bluetooth beacons to detect [your presence] $_{i_p}$ at affiliated venues.	$\sqsupseteq$

Table 1: Examples of evidence frames, e , and the subsumptive relationships we parse between question and answer.

Privacy questions may include terms different from those used in a policy. These might be simple taxonomic relationships such as a term used in the question being a hyponym/hypernym of terms used in a privacy policy. Yet other times the relationships may be more complicated. Sometimes regulations offer guidance about the way in which these terms are to be interpreted. Often they do not. To automatically answer privacy questions, it is critical to have clear and consistent definitions of relationships between terms, and to be able to spell out assumptions that the system might be making in answering questions. In this work, we consider 5 different possible relationships between terms used in a user question and information discussed in a relevant evidence frame. We will use i_p to denote an information type in a given privacy policy and i_q to refer to the information type in a user query. Similar relationships can be defined for terms referring to data practices such as the "collection", "sharing", "retention" or "deletion" of different data types.

Definition 2 (Hyponym). We define $i_p \subsetneq i_q$, if i_p is a more specific instantiation of i_q . For example, $i_p \subsetneq i_q$ for i_p ="GPS information", i_q ="location".

Definition 3 (Hypernym). We define $i_p \supsetneq i_q$, if i_q is a more specific instantiation of i_p . For example, $i_p \supsetneq i_q$ for i_p ="personal information", i_q ="location" under the Children Online Privacy Protection Act (COPPA).

Definition 4 (Synonym). We define i_p as a synonym or paraphrase of i_q , or $i_p \equiv i_q$, if i_p is another way of referring to i_q , but is neither a subset nor a superset. For example, $i_p \equiv i_q$ for i_p ="personal information", i_q ="personal data".

Definition 5 (Pseudo-hyponym). We define i_p as a pseudo-hyponym of i_q , or $i_p \sqsubseteq i_q$, if i_q could subsume i_p , but there is insufficient information to establish the subsumption. For example, $i_p \sqsubseteq i_q$ for i_p ="information about how you get to work", i_q ="health information", since "information about how you get to work" could be health information if it includes activities like walking or biking.

Definition 6 (Pseudo-hypernym). We define i_p as a pseudo-hypernym of i_q , or $i_p \sqsupseteq i_q$, if i_p could subsume i_q , but there is insufficient information to establish this. For example, $i_p \sqsupseteq i_q$ for i_p ="your presence at affiliated venues based on bluetooth beacons", i_q ="information about (your) visits to abortion clinics", since visits could be captured based on bluetooth beacons at a clinic.

We define these subsumptive relationships for the fields in the evidence frame e (Table 1). Any of these fields may be unspecified in a given policy statement, or be a set of values if multiple values are specified in a statement.

4 Annotation Process

An objective of this study is to estimate the importance of taxonomic inferences and of knowledge of regulatory disclosure requirements in accurately answering privacy questions. To do so, we identify ground truth answers to the 7 prototypical privacy questions selected for this study and use these answers to conduct an evaluation of generative AI models. Specifically, our annotators were requested to provide annotations at two levels: (1) First, they were asked to read the policies and identify all relevant reference frames for each question. (2) For each policy-question pair, once they were done identifying relevant references frames, they were asked to provide an overall categorical answer to the particular question. Here, annotators were requested to choose one of the following options:

1. **Explicit positive statement:** The policy states that the app can engage in this practice. This includes positive statements with associated conditions such as *'we may collect your location information if [you turn on the location permission for the app]...'* ;
2. **Explicit negative statement:** The policy states that it does not engage in this practice;
3. **Implicit positive statement:** While not explicit, the policy implies that it could engage in this practice by including a positive statement that combines the practice mentioned in the question (or a hypernym thereof) with the data type used in the question (or a hypernym thereof) with the positive statement including at least a hypernym of the data practice or a hypernym of the data type;
4. **Silence:** The policy is silent about this practice.

5. **Contradictory statements:** The policy includes contradictory statements such an explicit positive statement and an explicit negative statement.
6. **Other:** This last option was made available just in case an annotator were to decide that none of the previous option were appropriate. This option was only selected twice, and manual review determined that in both cases the annotator should have selected "Silence" as their categorical answer and adjusted their responses accordingly.

Note that the above options intentionally ignore applicable disclosure requirements. Instead, we wanted annotations that would enable us to draw different conclusions depending on assumptions made about disclosure requirements. The annotation process took 2 months and involved weekly meetings with annotators to discuss their annotations. The process resulted in 749 categorical answers and supporting frame annotations. Out of the 210 policy-question pairs, 92 were annotated by 3 annotators, 117 by 4 annotators and 1 by 5 annotators.

5 Annotations Analysis

Answer Category 1	Answer Category 2	Proportion
Explicit Positive	Implicit Positive	47.89%
Contradiction	Implicit Positive	2.82%
Explicit Negative	Implicit Positive	7.04%
Contradiction	Explicit Negative	2.82%
Explicit Positive	Silence	9.86%
Implicit Positive	Silence	16.9%
Explicit Negative	Explicit Positive	4.23%
Contradiction	Explicit Positive	1.41%
Explicit Negative	Silence	7.04%

Table 2: Analysis of the 19 policy-question pairs for which there was no majority consensus on a categorical answer. The first two columns represent divergent answer categories selected by annotators. We consider all pairwise disagreements between annotators and report the proportion of these disagreements.

nized around the remaining 5 options: *explicit positive statement*, *explicit negative statement*, *implicit positive statement*, *silence*, and *contradictory statements*.

Answer Distributions Figure 1 displays the distributions of categorical answers provided by our expert annotators for each of the 7 questions in our corpus. For

We analyze the annotations we collected for the 7 privacy questions and 30 mobile app privacy policies considered in this study. We are interested in estimating the prevalence of situations that require taxonomic reasoning or require making assumptions about regulatory disclosure requirements. As mentioned in Section 7, while annotators were asked to select from 6 options, the option "Other" was only selected twice and in both instances was truly intended to mean "Silence". Accordingly, categorical answers are orga-

each question, Figure 1a gives equal weight to each policy-question-annotator tuple independently of whether some policy-question pairs received 3, 4, or 5 annotations. Figure 1b on the other hand gives equal weight to each policy-question pair, using the majority annotation for each policy-question pair when a majority annotation exists. When a policy-question pair did not yield a majority answer, it is counted as "no consensus". For instance in Figure 1a, we can see that our annotators found a number of the mobile apps explicitly disclose collecting location information (Q1 - "explicit positive"), whereas much fewer apps disclose accessing the phone's camera (Q2 - "explicit positive"). This can also be seen in Figure 1b, where we can see that for many of the 30 apps, a majority of the annotators reported that the app can collect location information. We can also see however that for a little over 10% of the apps the annotators could not agree on whether or not the app can collect the user's location information ("no consensus"). At first glance, the fact that over 80 percent of the 30 mobile apps were determined by a majority of annotators to be allowed to collect location information may seem high. It should be noted however that "explicit positive" includes positive statements about hyponyms. In the case of location information, our expert annotators decided that terms such as "IP address" qualified as hyponyms of location. Further, annotators do not always converge on the same answer and, despite an annotation process that involved providing annotators with detailed instructions and weekly meetings where they were asked to discuss cases with which they struggled, a number of policy-question pairs did not admit a majority answer, with Q5, the question about whether an app can share the user's location with advertisers, leading to the highest level of disagreement among annotators - about a third of the policy-question pairs. This finding is nothing new and others have reported similar difficulty in getting annotators, including expert annotators with a legal background, to agree on privacy policy annotations (e.g., [13]). It is also worth noting that, while on a few occasions, an annotator reached the conclusion that a policy seemed to have contradictory statements about a particular question (see Figure 1a), this option was never selected by a majority of annotators for any of the policy-question pairs (see Figure 1b). For instance, for the privacy policy of 'Future Self', one annotator identified that the policy stated 'We DO NOT collect, store or use any personal information while you visit, download or upgrade our Applications', yet in another place 'We may use personal information submitted by you only for the following purposes: Help us develop, deliver, and improve our products and services and supply higher quality service'. This annotator reported this as a contradiction but was the only one to do so.

We computed the agreement between annotators on policy-question pair annotations by randomly sampling three annotations for that particular policy-question pair, producing a Fleiss' κ of 52.15 on these categorical answers, representing moderate agreement, which, as already noted earlier, is consistent with earlier findings that even expert annotators often disagree on the interpretation of privacy policy texts [13]. Analysis of 19 policy-question pairs for which there was no majority consensus on a categorical answer (see Table 2) indicates that

the most frequent source of disagreement (nearly half the disagreements) had to do with annotators being split between "explicit positive" and "implicit positive". For example, the privacy policy of the 'Hotels' app contains the following statements: "When you use our platform , Apps , or associated tools or services , we may collect the following kinds of personal information from you as needed : * Name , ... , and home , business , and billing addresses (including street and postal code) ...; * Geolocation ", and "Your personal information may be shared to help you book your travel and / or vacation ..."

While some annotators interpreted this as an *explicit positive statement* (the app can share the user's location), other annotators interpreted this differently, considering that the statement only describes sharing 'personal information', only implying that the app *could* share location information with other entities. This example illustrates how tenuous some sources of disagreement can be. As will be further detailed below, we sometimes opted to abstract away these finer sources of disagreement and focus on coarser interpretations of the annotations when these coarser annotations are sufficient to answer a given privacy question.

1.1 *Refers to regulation*: Do generative models refer to regulation when providing answers to privacy questions?

Below is the privacy policy of an app I am thinking of downloading on my smartphone. Can you tell whether this app could collect my <location information>. Please answer by selecting one of the following options: (1) Yes, (2) No, (3) It depends, (4) Other. In no more than 3 sentences, justify your answer. In particular, if you select "(3) It depends", please explain what the answer depends on.

1.2: *Assume disclosure requirements*

Assume that applicable regulations require that if the app can collect the user's location, this has to be disclosed in the policy. Below is the privacy policy of an app I am thinking of downloading on my smartphone. Can you tell whether this app could collect my location information. Please answer by selecting one of the following options: (1) Yes, (2) No, (3) It depends, (4) Other. In no more than 3 sentences, justify your answer. In particular, if you select "(3) It depends", please explain what the answer depends on.

1.3: *Assume no disclosure requirements*

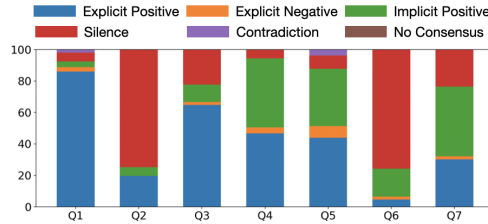
Assume that applicable regulations do NOT require an app to disclose the collection of location information. Below is the privacy policy of an app I am thinking of downloading on my smartphone. Can you tell whether this app could collect my location information. Please answer by selecting one of the following options: (1) Yes, (2) No, (3) It depends, (4) Other. In no more than 3 sentences, justify your answer. In particular, if you select "(3) It depends", please explain what the answer depends on.

1.4: *Identify categorical answer*

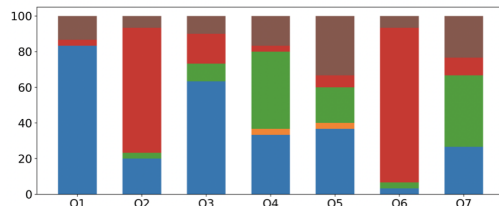
Below is the privacy policy of an app I am thinking of downloading on my smartphone. Does this policy indicate that the app can collect my location information? First give an answer and then state supporting evidence. The answer must be one of the following five options: (a) It explicitly indicates it can collect your location information, (b) It implies but does not explicitly state that it can collect your location information, (c) it explicitly indicates it will not collect your location information, (d) the policy does not indicate one way or the other whether it can collect your location information, or (e) this policy includes seemingly contradictory statements.

Table 3: Prompting Strategies for generative models.

Importance of Taxonomic Inferences We are interested in understanding the extent to which answering privacy questions requires making taxonomic inferences over the text of privacy policies. We start by measuring the proportion of instances where annotators conclude that a policy includes an explicit statement that the app engages in a practice, but did not identify an evidence frame that included an exact match for the information type and practice in the question. For instance, when we analyzed annotations for question Q1 where annotators labeled Q1 as "*explicit positive*". As we focus on these instances, we can further distinguish between instances where the explicit match relies on an evidence frame with an exact match for the data practice and data type, and instances where there is no exact match. The latter instances can further be subdivided into instances that only include evidence frames with hyponyms (e.g., hyponym of the data practice and/or data type) and instances that include a combination of labels other than exact matches.



(a) All annotations



(b) Aggregated annotations (majority vote)

Fig. 1: Distribution of categorical answers provided by expert annotators for each question. Fig(a) represents all annotations provided by the annotators, Fig(b) is the answer provided by majority of annotators for a policy-question pair. If not available, the answer is marked as ‘No consensus’.

We find that a considerable proportion (25.9%) of instances require making taxonomic inferences over the collection of location information. Out of 27 sets of annotations where the annotator reported explicit collection of location information, 2 contained only hyponyms, and 5 had evidence frames that included a combination of hyponyms, hypernyms, pseudo-hyponyms, and pseudo-hypernyms. For those two policies with evidence frames that only included hyponyms the most prevalent hyponyms that were reported were "geolocation", "GPS", and "precise location". For the five policies with evidence frames containing a mix of relationship types, we find a relatively equal distribution across hyponyms, hypernyms, and pseudo-hypernyms. There were however far fewer instances of evidence frames with pseudo-hyponyms. Most of the hyponyms labeled in these cases were also found in the cases where only hyponyms were la-

beled ("precise location", "GPS", "geolocation"). One notable example of a hy-

ponym that was common in the mixed annotations but not in the hyponym-only instances was "IP address". Other less common labeled as hyponyms included "Bluetooth", "MAC address", "cellular network data", "billing address", "zip code", "delivery options", "nearby cell towers", and "region information". The diversity of these terms illustrates how challenging answering privacy questions can be.

Mixed annotation cases included terms labeled as hypernyms, pseudo-hypernyms, and pseudo-hyponyms. Common terms were "contact information", "personal information", and other variants of "user information" or "user data" (commonly labeled as both hypernyms and pseudo-hypernyms). Other terms in the annotations were "travel information" (hypernym), "digital payment information" (pseudo-hypernym⁶), "unique identifiers" (pseudo-hypernym), "usage data" (pseudo-hypernym), "advertising identifiers" (pseudo-hypernym), "technical data" (pseudo-hypernym), "sensory and motion data" (pseudo-hypernym), and "driving event data" (pseudo-hypernym). Some of these choices are obviously subjective and help explain disagreements among annotators. Though much less common, notable terms labeled as pseudo-hyponyms included "zip code" and "IP address". In summary, the above analysis suggests that taxonomic reasoning is commonly required when it comes to answering privacy questions.

The Importance of Interpreting Policy Silence Our second objective is to estimate the prevalence of policy silence. Interpretation of silence varies based on applicable regulatory frameworks, and the disclosure requirements specified in regulation. As reported in Figure 1, policy silence turns out to be common. Among the 210 policy-question pairs considered in our study, 26.67% yielded majority annotations indicating "Silence" with some questions such as Q2 and Q6 yielding such answers for well over half of the mobile app policies. This suggests privacy QA systems need the ability to recognize silence, as well as the ability to interpret silence in light of applicable regulatory disclosure requirements.

6 Answering Questions with Generative Large Language Models

We evaluate the ability of generative large language models to answer privacy questions. We examine: (1) Do generative models refer to regulation when providing answers to privacy questions?, (2) When prompted to consider regulatory disclosure requirements, can generative models correctly interpret policies to answer privacy questions?, (3) Can generative models successfully identify scenarios from a policy that are applicable to each question?

⁶ Our interpretation of this pseudo-hypernym is that payment information can include Point-of-Sale data and therefore be indicative of a location visited by the user.

Model Name	Accuracy	Silence (P/R/F1)
Majority	33.33	0/0/0
Llama-2-70B-chat	17.62	16.67/5.36/8.11
gpt-3.5-turbo-16k	38.57	30.68/96.43/46.55
gpt-4-1106-preview	66.67	43.9/64.28/52.17

Table 4: Model classification results. The middle column reports accuracy across all 5 options in Prompt 1.4 (a) - (e). (P/R/F1) represents the precision, recall, and F1 scores for the ‘Silence’ category - category (d) in Prompt 1.4. If the ground truth does not have consensus, we reward the language models if they align with any expert annotator’s answers.

cable, or indicate that the answer to the privacy question may be influenced by relevant regulation. We analyze GPT-4 responses for all 210 policy-question pairs, and find no mention of regulation or clarification questions about the user’s residence. We further qualitatively analyze 50 of these responses to confirm that mentions of regulation are completely absent in generative model responses.

When prompted to consider regulatory disclosure requirements, can generative models correctly interpret policies to answer privacy questions? In order to understand whether generative models can take into account applicable privacy regulation when nudged to do so, we prompt GPT-4 with prompts 1.2 and 1.3 as described in Table 3; note that the prompts in the Table use collection of location as an example. Together these two prompts explore two complementary scenarios. Prompt 1.2 asks the model to assume that regulation requires an app to disclose the queried data practice if the app engages in it, whereas prompt 1.3 specifies that applicable regulation does not require disclosing the data practice. If a policy is silent about a practice and disclosure of the practice is required, the company is not allowed to engage in that practice. In contrast, if disclosure is not required and the policy is silent, the company could engage in the practice. Analysis of responses produced by GPT-4 show the following: (1) For those cases wherein the policy is silent and regulation requires disclosure (Prompt 1.2), GPT-4 correctly responds that the app cannot engage in the data practice only in 39.3% of the time, (2) For those cases wherein the policy is silent and regulation does not require disclosure (Prompt 1.3), GPT-4 correctly responds that the app can engage in the data practice *only 5.36%* of the time, (3) Analysis comparing model answers to prompt 1.2 and 1.3 for the same policy-question pair (which only differ on whether applicable regulation requires disclosure of the data practice or not) indicates that, on the whole, there is little

Do generative models refer to regulation when answering privacy questions? In order to understand if generative models make reference to applicable privacy regulation when formulating answers to privacy questions, we prompt GPT-4⁷, a current state-of-the-art generative language model with prompt 1.1 as described in Table 3. Ideally, a model that references applicable regulation would either clarify with the user which regulatory regime may be appli-

⁷ Accessed Nov 30, 2023.

model sensitivity to disclosure requirements even when explicitly specified (as is the case in prompt 1.2 and prompt 1.3). Even when the policy is silent, 51.8% of the time the model does not change its answer as the assumptions about disclosure requirements are changed in the prompts.

Can generative models accurately interpret privacy policies? We evaluate the ability of models to accurately interpret a privacy policy and identify the categorical answer for a given question. As shown previously, this can often require making taxonomic inferences over the content of privacy policy text. We evaluate the performance of Llama-2-70B-chat [15]⁸, GPT-4 [11], and GPT-3.5-turbo (Table 4). We also evaluate the ability of models to accurately recognize policy silence. We find GPT-4 is the best performing model, achieving nearly 30 points in accuracy over a majority baseline at identifying answers to questions. We also find it is able to recognize silence with a performance of 52.17 F1, and conceivably could be used to recognize when policies are silent. The QA system could then rely on a post-processing step that interprets silence based on applicable regulatory disclosure requirements.

In addition we find evidence that GPT-4 can successfully make several kinds of taxonomic inferences that are necessary to construct answers for privacy questions though further investigation is required. For example, for the question ‘*Q1: Does this app collect my location*’ and the privacy policy of Twitch which contains the statement ‘*Examples of such information we automatically collect include Internet Protocol address (“IP Address”), a unique user ID, device and browser types and identifiers, referring and exit page addresses, software and system type, and information about your usage of Twitch Services.*’, GPT-4 correctly infers that IP addresses could be construed as a hyponym of location and states that ‘Collecting an IP address can be used to determine an approximate location of a user, which constitutes location information.’

Can explicit taxonomic information improve generative model performance? Finally, we examine whether incorporating explicit taxonomic knowledge and prompting the generative AI model to differentiate between explicit and implicit statements helps enhance performance. To this end, we experiment with two approaches: (1) providing a handcrafted taxonomy of relevant terms constructed by domain experts along with a prompt designed to support the model in taking advantage of this taxonomic knowledge by differentiating between implicit and explicit statements, and (2) a baseline prompt that does not include taxonomic knowledge and does not differentiate between implicit and explicit statements. Here for the purpose of comparing performance, because the baseline prompt does not distinguish between implicit and explicit statement, accuracy is simply organized around four possible answers for each privacy question: (a) The policy indicates it can engage in the data practice, (b) the policy states it will not engage in the data practice, (c) the policy does not indicate one way or

⁸ We find that Llama-2 often generates responses we are unable to map to any of our categories, in these cases we do not assign a positive score to the model’s prediction.

the other whether it can engage in the data practice, or (d) this policy includes seemingly contradictory statements. Performance of the prompts that include taxonomic knowledge is computed by bundling implicit statements and explicit statements to allow for comparison with the baseline. In the results reported here, we considered prompts built using two differences sources of taxonomic knowledge: (1) taxonomic knowledge manually constructed by domain experts in the context of the MAPS privacy policy compliance system [19], (2) a collection of hyponyms and hypernyms automatically mined from the text of 30 privacy policies as detailed in [3].

As described in Table 5, we find explicitly describing information types can considerably improve GPT-4 performance, suggesting that generative models do not completely capture the taxonomic reasoning necessary to process the content of privacy policies. While examining the source of this improvement, we find that performance is most improved for questions concerning ‘credit card information’ (Q3;Q7), which were challenging for baseline generative models and is reflected in Figure 2. Figure 2 also illustrates that in one case (Q2), including taxonomic information causes a decrease in performance. This is likely due to neither taxonomy including information for terms related to "camera", and rather only include taxonomic information for "access". When examining the predicted answers against the majority answers from annotators, we notice that GPT-4 with the added taxonomic information for "access" over-predicts positive instances of camera access (majority annotation/gold-label: 7, GPT-4 baseline: 14, GPT-4 + Extracted hyponyms/hypernyms [4]: 18, GPT-4 + MAPS [19]: 15). We hypothesize that the performance would be improved by having a more comprehensive taxonomy for "camera access", which would reduce the instances of false positives.

Model Name	Acc.	Positive (P/R/F1)	Silence (P/R/F1)
GPT-4	76.19	78.63/73.60/76.03	48.35/78.57/59.86
GPT-4 + Extracted [4]	76.67	57.57/91.35/70.63	53.34/42.86/47.53
GPT-4 + MAPS [19]	82.86	57.57/91.35/70.63	60.29/73.21/66.13

Table 5: Model classification results, with and without taxonomic information. The middle column reports accuracy across 4 options: (a) The policy indicates it can engage in the data practice, (b) the policy states it will not engage in the data practice, (c) the policy does not indicate one way or the other whether it can engage in the data practice, or (d) this policy includes seemingly contradictory statements. (P/R/F1) represents the precision, recall, and F1 scores for categories (a) and (c). Models do not predict (b) or (d) for any sample in our dataset, hence we omit their discussion here.

7 Discussion and Future Work

Our results provide evidence that effective privacy QA functionality needs to support taxonomic reasoning, given the diversity of terms in privacy policies. Many common privacy questions, even seemingly simple questions, clearly require this type of reasoning. Our study also illustrates the prevalence of silence in privacy policies, highlighting the need for systems capable of reasoning about regulatory disclosure requirements.

We acknowledge the scope of the study was limited to a small set of mobile apps and questions. The scale was constrained by the time and effort required to collect detailed annotations. We strove to identify a somewhat representative collection of apps by identifying both popular and less popular apps within six categories of mobile apps. We hope this intensive effort sheds light on the capabilities required to answer privacy questions, and provide impetus for future work to develop precise QA systems.

An objective of our study was to assess the ability of popular generative AI tools to answer privacy questions. While generative AI tools have recorded tremendous progress in recent years, our study suggests that, in their present form, these tools do not adequately make many of the taxonomic inferences required to accurately answer privacy questions. In addition, when unprompted, these tools seem to be unaware of the impact of regulatory disclosure requirements on interpreting silence in the text of privacy policies, which further limits their ability to generate accurate answers to common privacy questions. Our research suggests however that with adequate prompt engineering such as lists of examples of relevant hyponyms, hypernoms, pseudonyms the performance of these tools can be improved. Clearly more work is required in this area and the results presented in this study have to be viewed as preliminary. Future research should explore opportunities to develop finer models, including the possibility of explicitly training specialized models on gazetteers of hyponyms, hypernoms, pseudonyms and even (within limits) lists of pseudo-hypernoms and pseudo-hyponyms of common terms found in the text of privacy policies.

In the short-term, another possible approach might be to use generative AI tools as advanced classifiers, using prompt engineering to force them to select from a small number of options such as those considered in our study (explicit positive, explicit negative, implicit positive, silence, contradictory). Results from

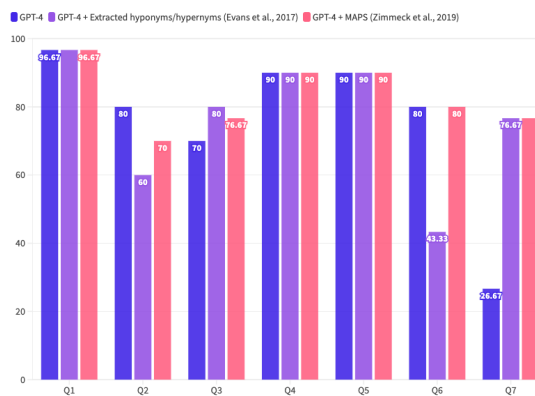


Fig. 2: GPT-4 performance, with and without taxonomic information, stratified by question type.

applying generative AI tools in this manner could then be post-processed to provide more complete answers. For instance, if "silence" is returned as the answer, a post-processing step could then be invoked to automatically interpret silence in light of applicable regulatory disclosure requirements. The resulting answer could then read: *"This privacy policy does not indicate whether or not this app shares your location with advertisers. Under California law, this app, which has over a million downloads, is required to disclose that it shares user location data with advertisers, if it does. Accordingly, this app probably does not share your location with advertisers, assuming that its policy is accurate."* In fact, an advantage of forcing generative AI tools to effectively act as classifiers is that one could then control the exact language of answers. This could simply be done by instantiating canned answers corresponding to different situations (e.g., silence about the sharing of one's location information in the above case).

Further, research has shown that users often struggle to even articulate their privacy questions. Rather than assuming that users are able to pose well-formed privacy question, one would ultimately want to develop privacy QA functionality that is capable of engaging in dialogues with users and through one or more iterations elicit the specific issue for which the user is seeking information.

The research presented herein is one step towards the development of practical privacy assistants capable of reliably answering people's everyday privacy questions. Ultimately we would want these assistants to be personalized and their answers to adapt to their particular users, including their level of education, technical expertise but also possibly what it takes to motivate them to pay attention to privacy risks and heed advice that might be provided in an answer. Related research on generating effective answers to common cybersecurity questions is discussed in a sister paper at this conference [\[2\]](#).

8 Conclusion

We document the significance of two different types of reasoning necessary to develop more accurate privacy question-answering systems. The first is the need to support taxonomic reasoning about related terms commonly found in the text of privacy policies. The second is the need to reason about regulatory disclosure requirements, given the prevalence of silence in privacy policy texts. Through a case study of 749 expert annotations of policies, we document the need for taxonomic reasoning and reasoning about regulatory data practice requirements. We additionally evaluate to what extent popular generative AI tools are able to reliably handle this type of reasoning and explore different ways of configuring them to do so. Ultimately we hope to facilitate the development of more powerful privacy question-answering systems capable of taking into account the particular regulatory provisions that apply to individual users (e.g., particular jurisdiction, whether the user is a child or an adult).

Acknowledgments. This research has been supported in part by grants from the National Science Foundation under the SaTC program (grants CNS-1914486, 1914444,

1914446) and under the REU program, the latter in part through CMU’s RE-USE Program (NSF grant 2150217).

References

1. Andow, B., et al.: {PolicyLint}: investigating internal privacy policy contradictions on google play. In: 28th USENIX security symposium (2019)
2. Balaji, A., Duesterwald, L., Yang, I., Priyanshu, A., Alfieri, C., Sadeh, N.: Generating effective answers to people’s everyday cybersecurity questions: An initial study. In: International Conference on Web Information Systems Engineering (2024)
3. Bhatia, J., et al.: Automated extraction of regulated information types using hyponymy relations. In: 2016 IEEE 24th International Requirements Engineering Conference Workshops. pp. 19–25. IEEE (2016)
4. Evans, M.C., et al.: An evaluation of constituency-based hyponymy extraction from privacy policies. In: IEEE 25th International Requirements Engineering Conference (2017)
5. Harkous, H., et al.: Polisis: Automated analysis and presentation of privacy policies using deep learning. arXiv preprint arXiv:1802.02561 (2018)
6. Kelley, P.G., et al.: A nutrition label for privacy. In: Proceedings of the 5th Symposium on Usable Privacy and Security. ACM (2009)
7. Kelley, P.G., et al.: Privacy as Part of the App Decision-Making Process. Association for Computing Machinery (2013)
8. Manandhar, S., et al.: Smart home privacy policies demystified: A study of availability, content, and coverage. In: 31st USENIX Security Symposium (2022)
9. Mysore Sathyendra, K., Wilson, S., Schaub, F., Zimmeck, S., Sadeh, N.: Identifying the provision of choices in privacy policy text. In: EMNLP (2017)
10. Oltramari, A., et al.: Privonto: A semantic framework for the analysis of privacy policies. Semantic Web (2017)
11. OpenAI: Gpt-4 technical report (2023)
12. Ravichander, A., et al.: Question answering for privacy policies: Combining computational and legal perspectives. In: EMNLP (Nov 2019)
13. Reidenberg, J.R., et al.: Disagreeable privacy policies: Mismatches between meaning and users’ understanding. Berkeley Tech. LJ **30**, 39 (2015)
14. Sadeh, N., Acquisti, A., Breaux, T., Cranor, L., McDonald, A., Reidenberg, J., Smith, N., Liu, F., Russell, N., Schaub, F., Wilson, S.: The usable privacy policy project: Combining crowdsourcing, machine learning and natural language processing to semi-automatically answer those privacy questions users care about. Tech. Rep. CMU-ISR-13-119, Carnegie Mellon University, Pittsburgh, Pennsylvania (December 2013)
15. Touvron, H., et al.: Llama 2: Open foundation and fine-tuned chat models (2023)
16. Wilson, S., et al.: 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) (2016)
17. Zhang, S., et al.: How usable are ios app privacy labels. Proc. Priv. Enhancing Technol. **2022**(4) (2022)
18. Zimmeck, S., Story, P., Smullen, D., Ravichander, A., Wang, Z., Reidenberg, J.R., Russell, N., Sadeh, N.: Maps: Scaling privacy compliance analysis to a million apps. Proceedings on Privacy Enhancing Technologies **2019**, 66 – 86 (2019)
19. Zimmeck, S., et al.: Maps: Scaling privacy compliance analysis to a million apps. Proceedings on Privacy Enhancing Technologies **2019**(3), 66–86 (2019)