Data Retention Disclosures in the Google Play Store: Opacity Remains the Norm

David Rodríguez*, Celia Fernández-Aller†, Jose M. Del Alamo*, Norman Sadeh‡
*ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain
{david.rtorrado, jm.delalamo}@upm.es
†ETSI Sistemas Informáticos, Universidad Politécnica de Madrid, Madrid, Spain
mariacelia.fernandez@upm.es
‡School of Computer Science, Carnegie Mellon University, Pittsburgh, USA
sadeh@cs.cmu.edu

***Abstract*—Privacy policies serve as the primary channel through which users are informed about the handling of their personal data, as required by regulations such as the General Data Protection Regulation (GDPR). This paper presents an evaluation of Android applications' privacy policies, focusing on how they articulate and disclose data retention periods. In this paper, we introduce a systematic approach that leverages Large Language Models to evaluate GDPR compliance regarding data retention disclosure across a diverse sample of 2,235 apps, demonstrating the applicability of the method at scale. Our approach reports a 0.904 F1 score, validated with a ground truth dataset manually annotated by legal experts and publicly released. Results show that over half of the examined policies are potentially non-compliant, with a significant subset indicating indefinite data retention and a high ratio of overlapping retention periods on the same privacy policy. This lack of compliance implies that those policies either fail to specify a retention period or provide unclear criteria for determining how long user data is kept. Thus, our study highlights the critical need to improve the clarity and enforcement of privacy policy practices, laying the groundwork for more transparent data governance.**

1. Introduction

In recent years, the technological landscape has undergone a dramatic transformation, underscored by the widespread adoption of smartphones and the explosion of mobile applications. These advancements, fueled by devices bristling with sensors, have catalyzed the vast collection and transmission of users' personal data across networks, leading to an era of unprecedented data accumulation. As the volume of stored data swells, so too do the privacy risks associated with its retention, ranging from data breaches to the misuse of aged information that could infringe upon individual privacy. The General Data Protection Regulation (GDPR) [1] addresses changing privacy needs by establishing clear limits on the collection and use of personal data.

In the context of rapid technological progress and increasing privacy concerns, this study examines the transparency of data retention disclosures in Android app privacy policies to ensure they meet GDPR standards for informing users about data handling and retention. Utilizing Large Language Models (LLMs), we analyzed 2,235 apps to assess GDPR compliance in their privacy policies. Our goal is to expose current practices, pinpoint transparency gaps, and highlight the importance of clear data retention disclosures.

2. Background and Related Work

GDPR emphasizes transparency as one of its seven key principles [6] upon which it is based, ensuring that individuals are informed. Among the specific mandates of this regulation is the requirement for clear disclosure of data retention periods in privacy policies, as detailed in Article 13(2)(b). It compels data controllers to transparently communicate either the specific duration for data storage or the criteria determining such periods, thus protecting user rights and facilitating informed decisions regarding their use of mobile applications.

Privacy policies serve as the primary conduit through which data controllers disclose their data handling practices to users, including data retention. Their inherent legal and technical jargon, combined with vague or overly broad statements, often obfuscates the true nature of data practices, leaving users ill-informed [9]. The difficulty in interpreting these policies can affect user comprehension and pose challenges for regulators and stakeholders in assessing compliance.

Given the challenges in understanding privacy policies, the pursuit of automation in privacy policy analysis has emerged as a promising solution, allowing evaluations at scale that provide a global perspective on personal data usage and retention practices. This technological approach can also enable Data Protection Agencies (DPAs) to save resources and enhance compliance monitoring efforts. Previous studies have demonstrated the applicability of automation in assessing legal compliance with respect to international transfers [5]. They used traditional Machine Learning methods based on pre-trained classifiers and observed a non-compliance ratio of over half of the apps. Thus, Machine Learning-based classifiers have been proven as an effective approach for privacy policy processing, specifically for the evaluation of compliance with other legal practices, such as the collection of personal data [14]. Recent studies have showcased the effectiveness of LLMs in analyzing and processing privacy policies [12]. Additionally, LLMs have been utilized to evaluate compliance with specific aspects such as personal data sharing, revealing that over 80% of apps did not meet the GDPR required standards [10].

Notably, related work [13] reveals that data retention, despite being one of the less frequently mentioned practices in privacy policies, occupies a substantial portion of the policy text when addressed. Another study [8] indicated that user preferences lean towards less intrusive data practices, including shorter data retention periods, underscoring the importance of our investigation into privacy policy transparency regarding data retention. To the best of our knowledge, our study is the first to systematically explore the disclosure of data retention periods in privacy policies based on an automated method and to demonstrate its effectiveness on a large-scale sample of privacy policies.

3. Method

This section presents our method, which leverages ChatGPT and its GPT-4 model [7] to assess the extent to which privacy policies comply with Article 13(2)(b) of the GDPR, regarding the disclosure of personal data retention periods. The method encompasses the creation of a ground-truth dataset, formulating the ChatGPT-based approach that identifies data retention periods, and validating it against the ground truth. The dataset used for this analysis, containing both the ground truth and validation results, has been publicly released [11] to contribute to the broader research community's efforts in scrutinizing privacy policy compliance with GDPR.

3.1. Ground truth

The foundation of our analysis rests on the OPP-115 dataset [13], widely recognized for its comprehensive annotations on privacy practices. This dataset is valuable for its explicit categorization of data retention practices, a focal point of our study. To tailor this dataset to our specific GDPR compliance assessment, two authors, including a senior data protection lawyer, manually reviewed each privacy policy segment that contained data retention statements as reported by OPP-115 annotators. This review process led to annotating six distinct categories that reflect the variance in GDPR compliance and transparency regarding data retention statements, as delineated in Table 1. These categories, labeled C0 through C5, establish the benchmarks for our evaluation of privacy policies, distinguishing between those that meet GDPR transparency requirements (C1-C5) and those that do not (C0).

Categories C1 and C2 represent policies with explicit disclosures of data retention periods, whether finite or indefinite. In contrast, categories C3, C4, and C5 encompass policies that define the retention period based on specific criteria or conditions rather than stating an explicit length of time. Although GDPR compliance for C5 necessitates additional scrutiny of its data retention purposes [4] (defined in Article 5(1)(b)), this study deems C5 valid without such analysis, acknowledging it goes beyond our current scope.

The dataset will be made publicly available upon the acceptance of this paper.

3.2. Method description

Our methodology employs the GPT-4 model, enhanced with the few-shot learning technique [3] and incorporates the case definitions outlined in Table 1 within the prompt. Additionally, we prompt the model whether the privacy policy expressly states no data retention, to exclude it from subsequent analyses. This approach enables a comprehensive classification of privacy policies according to various levels of transparency regarding data retention periods.

We have adjusted our approach to allow ChatGPT to identify and report multiple cases in which the privacy policy could be classified. This allows us to conduct a comprehensive study, identifying as many different data retention periods as the privacy policy declares. We encouraged the model to output its results in a Python list, aiding in the automated parsing of its outputs, a crucial feature for applying our method on a large scale.

3.3. Validation

Our method achieved outstanding results in evaluating GDPR compliance with data retention period disclosures. We measured an accuracy of 0.939, precision of 0.846, recall of 0.971, and an F1 score of 0.904. These metrics highlight the effectiveness of our methodology in accurately assessing privacy policies' compliance with GDPR's transparency mandates. Notably, the high recall rate indicates the method's proficiency in capturing instances of potential non-compliance. This serendipitous result reinforces the robustness of our approach, ensuring that few non-compliant policies are overlooked.

TABLE 1: Cases where privacy policies are categorized according to transparency and personal data retention periods.

Case	Description
C0	No data retention period is indicated in the privacy policy.
C1	A specific data retention period is indicated (e.g., days, weeks, months...).
C2	Indicate that the data will be stored indefinitely.
C3	A criterion is determined during which a defined period during which the data will be stored can be understood (e.g., as long as the user has an active account).
C4	It is indicated that personal data will be stored for an unspecified period, for fraud prevention, legal, or security reasons.
C5	It is indicated that personal data will be stored for an unspecified period, for purposes other than fraud prevention, legal, or security.

4. Evaluation in the wild

This section details the experiment conducted on a set of 2,235 Android applications. The proposed method has been utilized to assess these applications' compliance with

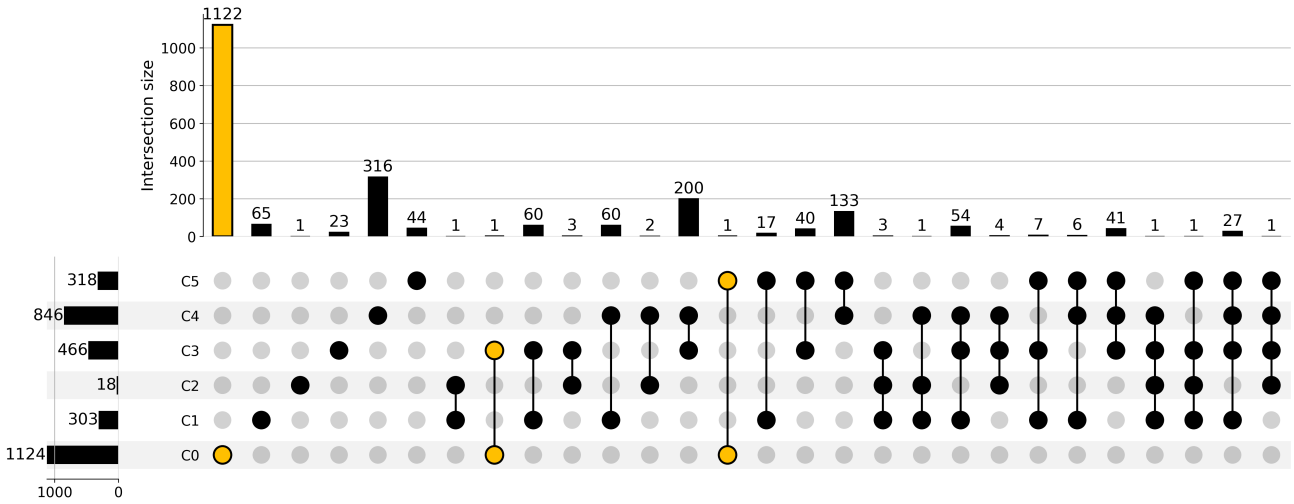


Figure 1: Number of privacy policies annotated according to each case (C0-C5). Dots represent those policies annotated with a single case (first six vertical bars) or multiple cases (from the seventh vertical bar onwards). Horizontal bars on the left side show the occurrence frequency of each case. The yellow color highlights potentially non-compliant policies. *Note: The two policies overlapping C0 and other cases should be considered wrongly annotated by ChatGPT.*

the transparency requirement of personal data retention periods mandated in the GDPR.

4.1. Experiment design and app selection

Our primary objective is to ascertain whether applications communicate their data retention periods transparently. This requires prior analysis of two conditions to determine that the application is retaining data and must declare the retention period(s): 1) the applicability of GDPR Article 13 to the application, inferred if the app processes personal data of EU subjects, and 2) the absence of a declaration in the privacy policy regarding non-retention of user data. The latter was rigorously evaluated through the method delineated in Section 3.2. To fulfill the first condition, we employed our dynamic app behavior evaluation platform.

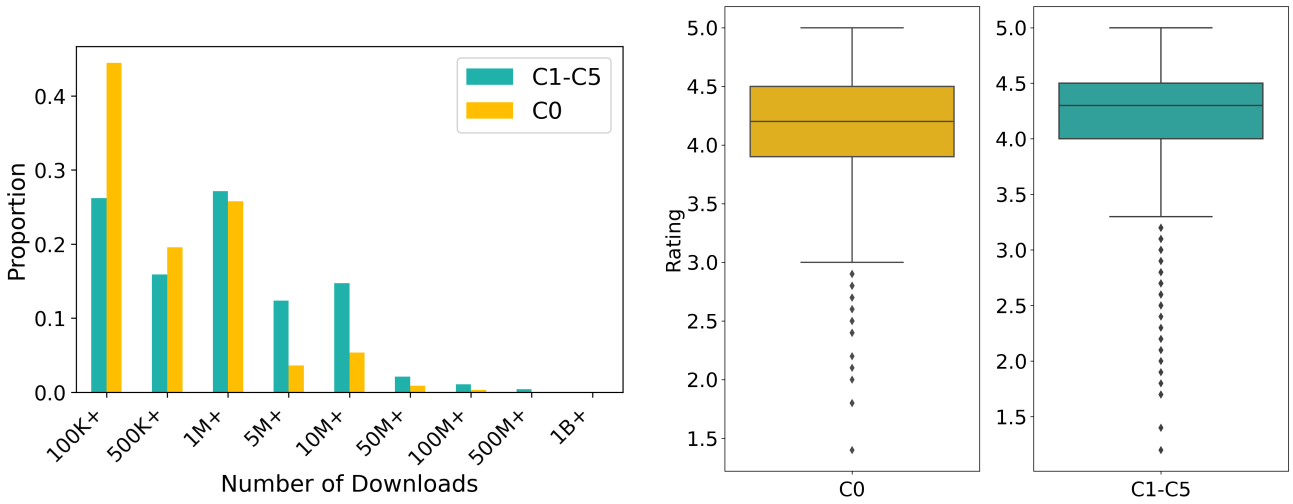
This platform employs a network of interconnected Docker containers designed to streamline the analysis process. Applications are systematically downloaded along with their respective privacy policies and stored for further analysis. Subsequently, these applications are installed on Xiaomi Redmi 10 devices to simulate real-user interaction, generating authentic network traffic. This traffic is captured via MiTM proxy and scrutinized to identify personal data transmissions. This allows us to determine if the applications meet criterion 1) outlined above and, therefore, must adhere to Art. 13(2)(b).

The selection process began with a dataset comprising over 4 million applications from the AndroZoo dataset [2]. To ensure diverse representation, we categorized these applications into three tiers based on their download counts as proxies for their popularity: from 100K to less than 1M for moderately popular apps, from 1M to less than 10M for highly popular apps, and 10M upwards for the most popular apps. From these tiers, we aimed to select a balanced sample of 10,000 applications, maintaining proportionality across the categories reflective of the original distribution, with 65.26% of the sample from the first tier, 27.40% from the second, and 7.33% from the third. These applications were subsequently downloaded and analyzed using a dynamic analysis platform. We successfully downloaded, installed, and executed 5,759 applications, intercepting their network connections. Of these, 3,478 applications, accounting for 60.39% of the subset, engaged in the transmission of personal data, and 2,307 had their privacy policies available in English, making them eligible for further examination under our study's criteria.

4.2. Results

In the analysis of 2,307 privacy policies, 72 applications were excluded due to a) 71 declaring no data storage and b) one providing no response, resulting in 2,235 policies for detailed evaluation. Our analysis divided these policies into two primary categories: those potentially non-compliant with GDPR (C0) and those likely compliant (C1-C5). Notably, a slight majority, 50.20%, were categorized as potentially non-compliant (C0), underscoring a significant challenge in meeting GDPR's transparency criteria for data retention periods. Figure 1 further highlights that data retention disclosures within privacy policies are unevenly distributed, with a predominant number of policies not specifying retention periods, and a considerable segment employing various criteria for determining data retention, indicative of the nuanced and often complex nature of privacy policy statements. This complexity and overlapping of criteria disclosed mirrors findings in the literature regarding the extension of such disclosures [13].

Our examination further explored the relationship between app popularity and GDPR compliance status. The analysis revealed a discernible pattern: applications categorized as potentially non-compliant (C0) generally exhibited lower download counts and marginally reduced user ratings compared to their counterparts in compliance categories C1-C5. Although a t-test indicated no significant differences in user ratings between the two groups



(a) Bar chart comparison between the app distribution of each set based on the number of downloads.

(b) Box plot comparison between the app distribution of each set based on the rating.

Figure 2: Comparison between the potential compliant (C1-C5) and non-compliant (C0) sets based on (a) the number of downloads and (b) the rating.

($p > 0.05$), a χ^2 (chi-squared) test on download counts yielded a statistically significant difference ($p < 0.05$). This observation strongly indicates that those responsible for the most popular applications are more likely to focus on legal and regulatory compliance.

Conclusion

This extensive analysis of Android applications has revealed a landscape where adherence to GDPR-mandated transparency in data retention disclosures is notably deficient, with over half of the evaluated policies potentially failing to meet the requirements. While 20.85% of policies explicitly allow for indefinite data retention, this study's focus on transparency acknowledges such declarations as compliant in that specific context, yet it underscores the broader privacy concerns they introduce. Moreover, the prevalence of policies detailing multiple data retention periods—34.18% of those analyzed—exemplifies the intricate nature of these documents, posing significant understanding difficulties for the average user.

Our findings advocate for enhanced regulatory oversight to ensure that privacy policies are compliant, clear, and accessible to users. In forthcoming research, we intend to broaden the scope of our analysis to encompass further aspects of transparency, examining how effectively users are informed about their rights concerning data erasure, modification, or access.

Acknowledgements

This work has been partially supported by the TED2021-130455A-I00 project funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU"/PRTR. This research has also been partially supported by the National Science Foundation under its Security and Trustworthy Computing Program (grant CNS-1914486).

References

[1] EUR-Lex—32016R0679 - EN, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[2] Kevin Allix, Tegawendé F. Bissyandé, Jacques Klein, and Yves Le Traon. Androzoo: Collecting millions of android apps for the research community. In *Proceedings of the 13th International Conference on Mining Software Repositories*, MSR '16, pages 468–471, New York, NY, USA, 2016. ACM.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[4] Bart van der Sloot Chris Jay Hoofnagle and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means*. *Information & Communications Technology Law*, 28(1):65–98, 2019.

[5] Danny S. Guamán, David Rodriguez, Jose M. del Alamo, and Jose Such. Automated gdpr compliance assessment for cross-border personal data transfers in android applications. *Computers Security*, 130:103262, 7 2023.

[6] Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: What it is and what it means. *Information and Communications Technology Law*, 28(1):65–98, Jan 2019.

[7] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, et al. Gpt-4 technical report. *arXiv*, March 2023. Accessed: 2024-03-27.

[8] Dimitris Potoglou, Fay Dunkerley, Sunil Patil, and Neil Robinson. Public preferences for internet surveillance, data retention and privacy enhancing services: Evidence from a pan-european study. *Computers in Human Behavior*, 75:811–825, 2017.

[9] Joel R. Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T. Graves, Fei Liu, Aleecia McDonald, Thomas B. Norton, Rohan Ramanath, N. Cameron Russell, Norman Sadeh, and Florian Schaub. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal*, 30(1):39–88, 2015.

[10] David Rodriguez, Jose M. Del Alamo, Celia Fernández-Aller, and Norman Sadeh. Sharing is not always caring: Delving into personal data transfer compliance in android apps. *IEEE Access*, 12:5256–5269, 2024.

[11] David Rodríguez, Celia Fernández, Jose M del Alamo, and Norman Sadeh. Data retention period disclosures in privacy policies, 2024. Publisher: Mendeley Data, Version: V1, doi: 10.17632/c4x958pzpm.1.

[12] Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, and Lei Fan. Policygpt: Automated analysis of privacy policies with large language models, 2023.

[13] 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. The creation and analysis of a website privacy policy corpus. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 3:1330–1340, 2016.

[14] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. Maps: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies*, 2019:66–86.