Disclosure Divergence: Measuring Privacy Policy and Data Safety Misalignment at Scale

Mst Eshita Khatun
Louisiana State University
Baton Rouge, Louisiana, USA
mkhatu3@lsu.edu

Lamine Noureddine
Louisiana State University
Baton Rouge, Louisiana, USA
lnoureddine@lsu.edu

Sideeq Bello
Louisiana State University
Baton Rouge, Louisiana, USA
sbell49@lsu.edu

Aisha Ali-Gombe
Louisiana State University
Baton Rouge, Louisiana, USA
aaligombe@lsu.edu

Abstract

With the rapid growth of mobile applications, user data privacy has become an increasing concern. While privacy policies describe how apps collect and share data, platforms such as Google Play provide Data Safety labels intended to summarize these practices. Because these disclosure channels are declared separately, they may present inconsistent representations of app data practices, creating uncertainty for users and regulators. In this work, we conducted a large-scale empirical study of disclosure consistency across 6,051 Android apps. Using an LLM-based extraction framework and a unified schema over 14 Google Play data categories and two operations (collection and sharing), we measure per-app and per-category consistency and introduce a sensitivity-weighted risk score that emphasizes high-risk data types. We find that misalignment disproportionately affects sensitive categories such as personal information and device identifiers, with sharing disclosures exhibiting lower consistency than collection disclosures. Elevated privac risk is concentrated in app categories associated with persistent monitoring and communication. Overall, our findings highlight structural gaps in current disclosure mechanisms and underscore the need for stronger verification and greater transparency in platform-level privacy reporting.

Keywords

Data Safety, User Privacy, Privacy Policy, Privacy Analysis, LLM

1 Introduction

The rapid growth of mobile applications has raised serious concerns about the privacy of user data [30, 32, 43, 62]. Mobile apps, in particular, have become deeply embedded in our everyday lives, continuously accessing and processing sensitive information such as location, contacts, identifiers, financial records, and health data. This pervasive data collection amplifies users' exposure to tracking, profiling, and secondary data uses that often occur without explicit consent or understanding. In response to these growing privacy risks, Google introduced Data Safety Labels on the Play Store in 2022 [20, 25], requiring developers to self-report what data their apps collect, share, and how such data are secured. Developers must also accompany these disclosures with privacy policies, which serve as formal, legally binding statements of an app's data-handling practices and responsibilities [9]. Together, Data Safety Labels and privacy policies represent the main transparency interface between developers and users, informing users about how their data are collected, shared, and secured, and are thus ostensibly designed to enhance user trust when downloading and using apps.

Despite these efforts, persistent uncertainty remains over whether these declarations accurately reflect actual data practices or merely offer a perception of transparency that fails to guarantee real protection of users' privacy data. In particular, the lack of consistency between what developers self-report in their Data Safety Labels and what they state in their privacy policies about user data collection and sharing may confuse users, weaken their trust, and compromise the reliability and efficiency of Google's overall privacy and transparency framework [28, 61].

Prior studies have largely examined these two transparency layers in isolation. Research on policy analysis (e.g., PolicyLint [7], PrivOnto [37], Polisis [22], Privacify [55], and PrivacyCheck [36]) has focused on identifying data practices, contradictions, and compliance gaps within privacy policies. Meanwhile, work on Data Safety and Privacy Labels (e.g., Zhang et al. [59–61], Khandelwal et al. [24, 25], Khedkar et al. [27] and Wang et al. [52]) has evaluated their usability, coverage, and reliability as self-reported disclosures. In terms of cross-layer analysis, two recent studies have examined inconsistencies between platform privacy labels and privacy policies in both Android and iOS ecosystems [4, 23]. These efforts provide important evidence that disclosure inconsistencies remain widespread across mobile platforms. However, several analytical dimensions remain underexplored. In particular, these prior works have given limited attention to how inconsistencies differ between data collection and data sharing practices, how disagreement patterns vary across data categories, and how the relative privacy sensitivity of different disclosure mismatches can be systematically characterized. As a result, while these existing studies establish the presence of disclosure inconsistencies, there remains a limited understanding of their structure, concentration, and potential privacy severity across the broader Google Play ecosystem.

Proceedings on Privacy Enhancing Technologies 2026(4), 213–231

https://doi.org/10.56553/popets-2026-0117

Our findings reveal widespread inconsistencies between privacy policies and Google Play Data Safety Labels. At the disclosure level, approximately one third of collection (33%) and sharing (31%) disclosures disagree across the two layers, while over 90% of apps contain at least one inconsistency in Data Safety Label disclosures. These discrepancies disproportionately affect sensitive categories such as Personal Information, Device Identifiers, Location, and Financial Data, and are primarily driven by omissions in Data Safety Labels relative to privacy policies. Importantly, an ablation analysis on generic sharing statements shows that the observed asymmetry between sharing and collection disclosures remains stable, reinforcing the robustness of our findings. The persistence of this asymmetry suggests that the discrepancies are not solely artifacts of broad policy language. Instead, they appear particularly pronounced in data-sharing disclosures and may partially reflect a service-provider loophole, where data practices described in privacy policies under third-party or service-provider contexts are not consistently represented in the corresponding Data Safety Labels. Moreover, our privacy risk scoring analysis shows that about half of the analyzed apps fall into a medium-risk tier and a small but non-trivial subset into a high-risk tier, indicating that disclosure misalignment remains pervasive even among highly rated and widely downloaded apps.

In summary, this work makes the following salient contributions:

- A large-scale measurement of 6,051 real-world Android apps spanning 14 Google Play data categories, quantifying inconsistencies in developers' reported data-collection and data-sharing practices across Data Safety Labels and Privacy Policies.
- A framework that systematically and rigorously quantifies the consistency of developers' privacy disclosures across Data Safety Labels and Privacy Policies, using a sensitivity-weighted risk scoring mechanism at both the app and data-category levels.
- Concrete insights and recommendations to help platform providers, regulators, auditors, and developers improve the accuracy and alignment of privacy disclosures, thereby enhancing user trust and engagement within the app ecosystem.
- To support reproducibility, we publicly release all artifacts associated with this study through a GitHub repository [26]

The remainder of this paper is organized as follows. Section 2 presents the background, motivation and research questions. Section 3 describes the methodology, while Section 4 reports the findings. Section 5 discusses additional qualitative analysis and key observations, recommendations, limitations, and future work. Section 6 reviews related work, and Section 7 concludes the paper.

2 Background & Motivation

Google Play's Data Safety section requires developers to self-report which categories of user data their apps collect and share, and for what purposes, as shown in Appendix A, Figure 11. Based on the developer documentation, the Google Play Data Safety schema defines 14 user data type categories—Location, Personal info, Financial info, Health and fitness, Messages, Photos and videos, Audio files, Files and docs, Calendar, Contacts, App activity, Web browsing, App info and performance, and Device or other IDs which together cover 38 specific data types such as Approximate location, Email address, Purchase history, Web browsing history, and Device or other IDs including others [20]. The label, which distinguishes between "data collected" and "data shared with third parties," is intended to give users a quick, standardized view of an app's data practices directly in the store interface. In parallel, developers are also required (in most jurisdictions by the law, such as GDPR [15] and CCPA [11]) to provide a privacy policy, typically hosted on an external website, that describes in more detail what data is processed, collected, and/or transferred; for which purposes; with which partners, and under what legal bases.

Together, these two artifacts form the primary interface for transparency between users and developers, where most users see the Data Safety label first and consult the privacy policy only if they have concerns. However, both artifacts are developer self-reported, and neither is systematically checked against the other or against the app's actual behavior at scale. It therefore remains unclear how closely these descriptions agree, how they differ across data categories, and what these differences imply for user privacy. Our study addresses this by quantifying the degree of alignment and sensitivity risk of misalignment between Data Safety Labels and privacy policies across a large corpus of Google Play apps.

2.1 Motivation

Mobile apps increasingly mediate sensitive aspects of daily life, from communication and social networking to banking, health, and mental well-being. When deciding whether to install an app, users must primarily rely on two sources of information about data practices: the Google Play Data Safety label and the app's privacy policy. If these disclosures are incomplete or inconsistent, users may receive an inaccurate or partial understanding of how their data is collected and shared, particularly concerning highly sensitive categories such as location, financial records, or health information.

This raises a practical question for both end users and platform governance: to what extent can current disclosure mechanisms be trusted to provide a consistent and reliable representation of an app's stated data practices? Therefore, our goal in this paper is to provide a large-scale, quantitative assessment of the alignment between Data Safety labels and privacy policies, and to characterize the severity of any mismatches using a sensitivity privacy risk score. By systematically measuring consistency (agreement between the two transparency layers), identifying the data types with the highest misalignment prevalence, and assessing the sensitivity risk across thousands of apps, we aim to strengthen the reliability and accountability of privacy transparency mechanisms on Google Play. This study will help users make informed decisions, enable regulators to evaluate compliance, and guide platform designers in improving disclosure mechanisms and auditing practices.

2.2 Research Questions

To address these gaps, this study investigates inconsistencies, misalignment prevalence and sensitivity risk between privacy policies and Google Play's Data Safety labels through the following research questions:

- RQ1. How consistent are privacy policies and Google Play's Data Safety labels when reporting data collection and sharing?
- RQ2. Which data categories are most frequently misaligned across the two disclosure layers?
- RQ3. How are apps distributed across low, medium, and high Sensitivity Risk Score tiers, and what does this distribution reveal about the severity of privacy misalignment between privacy policies and Data Safety labels?
- RQ4. Which app categories exhibit the highest privacy misalignment risk, and how is risk distributed within those categories?

Overall, these questions allow us to provide insight into how well privacy disclosures align in practice. By quantifying how often misalignment occurs, and which apps and data categories are most affected, our exploratory analysis highlights concrete risks for user privacy and helps guide more trustworthy transparency mechanisms.

3 Methodology

This section presents the methodology of our cross-layer privacy disclosure analysis pipeline for examining developers' privacy disclosures regarding user data across privacy policy documents and Data Safety Labels, as illustrated in Figure 1. Our methodology consists of two main phases: Phase I, data collection; Phase II, data modeling; followed by quantitative analysis and the derivation of insights related to user data privacy.

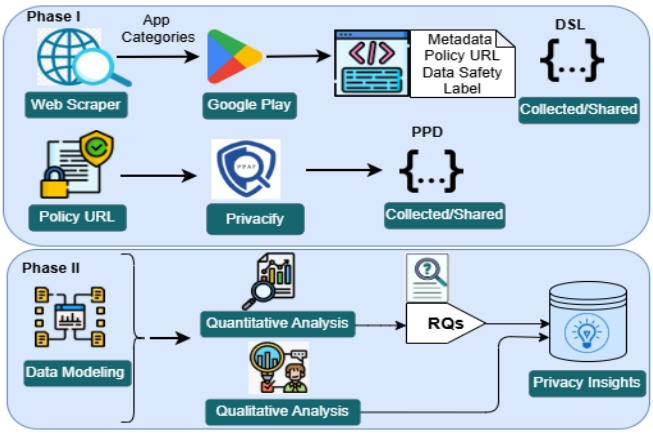

Figure 1: Cross-layer privacy disclosure analysis pipeline

3.1 Data Collection (Phase I)

To investigate inconsistencies, the prevalence of misalignment, and the sensitivity risk of privacy disclosure, we first created a dataset of mobile applications from the Google Play Store. The data collection phase involved gathering three primary artifacts for each selected app: (1) the app's privacy policy, (2) the Data Safety label displayed on its store listing, and (3) relevant metadata such as the app category, user ratings, and number of downloads. These artifacts serve as the foundational sources for evaluating our research questions. We developed a custom Python-based web-crawling toolkit targeting Google Play Store listings. The toolkit consists of modular scripts that (i) enumerate mobile applications from category-specific search queries, (ii) retrieve their Data Safety Labels, (iii) store the corresponding privacy policies, and (iv) collect basic app-level metadata (category, rating, reviews, downloads, and price). All components are implemented using `requests` and `BeautifulSoup4` for HTML fetching and parsing. Our crawler focuses on English-language apps, filtering out non-English listings during collection. This design yields a matched corpus of apps for which we have both Data Safety disclosures and privacy policies.

3.1.1 Data Safety Labels. We construct the Data Safety Label corpus starting from application category-specific search queries on the Google Play Store. For each query, our URL collection script crawls paginated search results, identifies individual app listing pages, and derives the corresponding Data Safety endpoints from their URLs. To avoid redundant work across runs, the crawler maintains a manifest of tuples of discovered Data Safety URLs and skips any URL that has already been processed. For each URL, a dedicated scraper fetches the Data Safety page and parses its HTML structure to locate the disclosure sections - *Data collected* and *Data shared*. It then iterates over the associated data categories, extracting both the sub-category label (data-type group) and its textual description. We normalize these disclosures into a structured JSON representation, where each app key maps to two sections (*Data collected* and *Data shared*), and each section is a mapping from a Data Safety category label (e.g., *Location*, *App activity*) to one or more concrete data-type strings (e.g., *Approximate location*, *App interactions*). For downstream analysis, we flatten this structure into a set of *(category, data-type)* pairs per app. Before processing a new app, the scraper checks this JSON store; if an app is already present, it is skipped. This incremental design allows the pipeline to be re-run safely and produces a stable, reproducible corpus of Data Safety label annotations for our analysis, with a representative JSON data structure illustrated in Appendix B listing 1.

3.1.2 Privacy Policy Documents. To obtain privacy policies that correspond to the same set of apps, we leverage the Data Safety endpoints as stable entry points into each app's detail page. A separate crawling component loads each app's listing (or Data Safety) URL, locates the outbound *Privacy policy* hyperlink in the app information panel, and records the resulting pair *(app name, policy URL)* in a manifest file. As with the Data Safety crawler, we deduplicate entries by both app identifier and URL, ensuring that previously seen policies are not revisited in later runs.

Given this manifest, a policy downloader fetches the content of each privacy policy, preferring English-language variants when multiple localized versions are available. The HTML is sanitized to remove boilerplate and navigation elements and converted into a text format that preserves structural cues such as headings, lists, and tables. The resulting normalized policies serve as input to our LLM-based policy analysis pipeline, from which we later extract app developer practices regarding data type collection and sharing claims for comparison against the Data Safety Labels.

3.1.3 Privacy Policy Analysis Pipeline. For the analysis of privacy policy documents, our methodology relies on an LLM-based system

to extract (i) the set of user data types an app claims to collect and (ii) the set of data types it reports sharing with third parties. To this end, we adopt and extend *Privacify* [55], a recent LLM-based privacy policy analysis system. Privacify implements a tokenization and chunking pipeline combined with an iterative map–reduce style summarization and recombination procedure, which enables processing long policies while preserving cross-section context. We further customize the final feature-extraction prompts so that the system outputs two sets of data types, C_{data} (collected) and S_{data} (shared with third parties), which we directly ingest as labels for our dataset construction.

Validation Methodology. To assess extraction accuracy across diverse and representative conditions, we constructed a validation set of 100 apps using a stratified sampling strategy that covers app categories, popularity, and policy complexity.

- For *app category coverage*, we selected 60 apps — approximately 5 per category — spanning 12 Google Play categories: Finance, Health and Fitness, Communication, Social, Shopping, Tools, Productivity, Education, Entertainment, Games, Lifestyle, and Personalization. This stratification ensures that category-specific vocabulary and disclosure patterns are represented in our evaluation, since privacy policy language varies substantially across domains.
- For *app popularity*, we selected 20 apps based on install and rating signals: 5 low-install apps (approximately 1K–5K downloads), 5 high-install apps (10B+ downloads), 5 highly rated apps (rating = 5.0), and 5 low-rated apps (rating $\leq$ 2.0).
- For *policy complexity*, we selected 20 apps based on structural and linguistic features: 5 very short policies (under 300 words), 5 very long policies (over 10,000 words), 5 policies with frequent third-party/SDK-related terms, and 5 policies using generic sharing language (e.g., "we may share," "partners," or "affiliates"). This strategy captures a range of policy structures and disclosure patterns to evaluate extraction performance under diverse privacy-policy conditions.

For each of the 100 apps, one author manually annotated whether the policy indicates collection and/or sharing of each of the 14 data categories. In many cases, third-party data sharing was described only in generic terms (e.g., "we may share your information"). When the policy did not specify which data categories were shared, we conservatively treated all data categories marked as collected for that app as shared with third parties. To assess the sensitivity of our data extraction results to this design choice, we further perform an ablation analysis under alternative treatments of generic sharing statements, described in Appendix G. To further evaluate annotation reliability, a subset of 40 apps was independently annotated by a second annotator, yielding Cohen's $\kappa = 0.78$ and 90% observed agreement. Disagreements were resolved through discussion before finalizing the ground truth labels.

LLM Backend Selection and Parameter Configuration. To assess whether extraction performance depends on the choice of LLM backend, we evaluated both the original Privacify backend, Mistral-7B, and our deployed backend, LLaMA 3.1 8B Instruct, on the same expanded 100-app validation set. For a fair comparison, both models were evaluated using the same preprocessing pipeline, chunking strategy, prompts, and output schema. Specifically, privacy policies were segmented using a token-based splitter with a chunk size of 3,000 tokens and an overlap of 100 tokens. Both models were run with deterministic decoding using temperature = 0.0, top-p = 1.0, and a maximum generation length of 6,000 tokens. The context window was configured to 8,500 tokens, and long policies were handled using chunking followed by map–reduce aggregation.

Table 1: LLM backend comparison on 100-app validation.

Operation	Model	Precision	Recall	F1-Score
Collected	LLaMA 3.1 8B	91.4%	93.0%	92.2%
	Mistral-7B	90.5%	89.5%	90.0%
Shared	LLaMA 3.1 8B	89.9%	92.1%	91.0%
	Mistral-7B	87.5%	88.1%	87.8%

Table 1 summarizes the extraction performance of the two LLM backends on the expanded 100-app validation set. LLaMA 3.1 8B Instruct consistently outperformed Mistral-7B across both collection and sharing extraction tasks. For collection disclosures, LLaMA 3.1 8B achieved an F1-score of 92.2% compared to 90.0% for Mistral-7B. The difference was more pronounced for sharing disclosures, where LLaMA 3.1 8B achieved an F1-score of 91.0% versus 87.8% for Mistral-7B. We also observed higher recall improvements for sharing extraction, suggesting that LLaMA 3.1 8B is more effective at identifying third-party sharing disclosures in privacy policies. Overall, these findings indicate that model choice has a measurable impact on extraction accuracy, particularly for sharing-related disclosures, while also demonstrating that the overall extraction pipeline remains stable across different LLM backends. Overall, these results indicate that LLM extraction is sufficiently accurate to support aggregate measurements and provides a scalable, reliable mechanism for generating consistently labeled data for privacy policy analysis. An example of the resulting JSON data structure is provided in Appendix C, Listing 2.

3.1.4 Dataset Overview. To construct the app corpus, we adopted a category-guided sampling strategy rather than relying on a single popularity ranking or top-chart list. Specifically, we defined category-specific search queries to cover a broad range of Google Play segments spanning diverse application domains and app genres. Apps were collected through query-based crawling across categories to better capture the heterogeneity of the Google Play ecosystem. For each query, our crawler retrieved app listing URLs from Google Play search results along with the corresponding Data Safety page, privacy-policy link, and available metadata.

Using this approach, we collected approximately 10K app entries from the Google Play Store between August and September 2025. We further removed duplicate records using the `APP Name` field and, when an app with a similar name appeared under multiple categories, we retained a single primary category. We also inspected the reported privacy policy URLs and discarded cases where the "policy" link actually pointed to unrelated websites (e.g., marketing pages or generic homepages) rather than a genuine privacy policy. This de-duplication and filtering step yielded a final corpus of 9,162

distinct applications spanning 419 application categories. These categories include generic genres such as *Games*, *Productivity*, and *Lifestyle* and privacy-sensitive verticals, such as *Health and Fitness*, *Medical*, *Mental health*, *Finance*, *Banking*, *Insurance*, *VPN apps*, *Messaging*, *Cloud storage*, *Password manager*, and *Kids education*. It also covers specialized app categories such as *IoT control*, *Smart home*, *Remote desktop* and *Wearables*. This range provides us with a diverse selection of applications, data types, and sensitivity levels to base our study on privacy disclosures. The summary of our dataset is shown in Table 2 including additional metrics such as free vs. paid , ratings, reviews and downloads.

Table 2: Summary of the app corpus used in our analysis.

Metric	Value
Number of apps	9,162
Number of categories	419
App Rating	4.19 (median 4.40)
Engagement (min-max)	5 – 212.4M (median 8.6K)
Apps with privacy policy	6,051 (66.0%)
Apps without privacy policy	3,111 (34.0%)
Free apps	8,757 (95.6%)
Paid apps	405 (4.4%)
Free apps with privacy policy	5,910 (64.5%)
Paid apps with privacy policy	141 (1.5%)
Free apps without privacy policy	2,847 (31.1%)
Paid apps without privacy policy	264 (2.9%)
Rating range (min–max)	1.0 – 5.0
Reviews range (min–max)	5 – 212.4M
Downloads range (min–max)	1K – 500M

3.2 Data Modeling (Phase II)

After collecting the disclosure artifacts and metadata, the next phase involved modeling and structuring the dataset to enable comparison across privacy policies and Data Safety labels. We transformed the raw disclosures into standardized representations by parsing and categorizing data practices into a unified schema aligned with Google Play's data categories and operations. This normalization process allows us to operationalize inconsistencies and misalignments at the data-type level and supports the computation of metrics such as misalignment prevalence and sensitivity-risk scores used in subsequent analysis.

We consider a set of apps indexed by $i = 1, \ldots, N$, and a set C of data categories (e.g., Location, Personal Info, Identifiers, Financial Info). For each app i, category $c \in C$, and operation $o \in \{\text{collect}, \text{share}\}$, we define binary indicators.

Binary Indicators

The Privacy Policy Document (PPD) indicator:

$$PPD_{i,c}(o) = \begin{cases} 1, & \text{if the privacy policy reports } c \text{ for } o, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The Data Safety Label (DSL) indicator:

$$DSL_{i,c}(o) = \begin{cases} 1, & \text{if the Data Safety label reports } c \text{ for } o, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Consistency in the Privacy Policy Document and Data Safety Label

We define consistency (alignment) as the agreement between the PPD and the DSL regarding whether a specific data category is collected or shared. This definition yields a binary consistency indicator for app i, category c, and operation o:

$$\text{Cons}_{i,c}(o) = \begin{cases} 1, & \text{if } PPD_{i,c}(o) = DSL_{i,c}(o) \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

where $\text{Cons}_{i,c}(o) = 1$ indicates a match (or an agreement) between the privacy policy and the Data Safety label for category c in operation o, and $\text{Cons}_{i,c}(o) = 0$ indicates a mismatch (misalignment or contradiction).

Consistency Score

And thus, the aggregated app-level consistency score for an app i and an operation o is given as:

$$\text{ConsScore}_i(o) = \frac{1}{T_C} \sum_{c \in C} \text{Cons}_{i,c}(o) \quad (4)$$

where C denotes the set of data categories and $T_C = |C|$ is the total number of categories.

Misalignment in the Privacy Policy

Misalignment in the Privacy Policy Document is when the PPD omits but the DSL reports operation on data category. This is given as:

$$U^{PPD}_{i,c}(o) = \begin{cases} 1, & \text{if } PPD_{i,c}(o) = 0 \text{ and } DSL_{i,c}(o) = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

Misalignment in the Data Safety Label

Misalignment in the Data Safety Label is when the DSL omits but the PPD reports operation on data category. This is given as:

$$U^{DSL}_{i,c}(o) = \begin{cases} 1, & \text{if } PPD_{i,c}(o) = 1 \text{ and } DSL_{i,c}(o) = 0, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Misalignment Score

Thus, the misalignment indicator for an app i, category c, and operation o is given as misalignment in the Privacy Policy *or* misalignment in the Data Safety Label.

$$\text{Mis}_{i,c}(o) = \begin{cases} 1, & \text{if } U^{PPD}_{i,c}(o) = 1 || \text{if } U^{DSL}_{i,c}(o) = 1 \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

where $\text{Mis}_{i,c}(o) = 1$ indicates a mismatch (contradiction or a disagreement) between the privacy policy and the Data Safety label for category c in operation o, and $\text{Mis}_{i,c}(o) = 0$ indicates a match (alignment).

And thus, the aggregated app-level misalignment score for an app i and an operation o is given as:

$$\text{MisScore}_i(o) = \frac{1}{T_C} \sum_{c \in C} \text{Mis}_{i,c}(o) \quad (8)$$

where C denotes the set of data categories and $T_C = |C|$ is the total number of categories.

Consistency Vs Misalignment Scores

Given the definition, Misalignment and Consistency form complementary measures over the unit interval, satisfying

$$\text{ConsScore}_i(o) + \text{MisScore}_i(o) = 1 \tag{9}$$

3.3 Cohen's Kappa Score

The consistency metrics defined in Equations 4 and 9 capture observed agreement between privacy policies and Google Play's Data Safety Labels; however, they do not adjust for agreement expected by chance. To address this limitation, we compute Cohen's kappa (κ), a chance-corrected measure of inter-rater agreement, separately for the two disclosure operations: *collection* and *sharing*. We treat the PPD and the DSL as two *raters*. For a fixed operation $o \in \{\text{collect}, \text{share}\}$, each *app–category pair* (i, c) constitutes one binary decision over the 14 Google data categories:

$$PPD_{i,c}(o) \in \{0, 1\}, \quad DSL_{i,c}(o) \in \{0, 1\},$$

where 1 indicates that the source discloses the category c under operation o, and 0 indicates it does not.

Let Ω_o denote the set of evaluated app–category pairs for operation o, and let $N_o = |\Omega_o|$. The observed agreement (P_o) is the fraction of pairs where the two sources match:

$$P_o(o) = \frac{1}{N_o} \sum_{(i,c)\in\Omega_o} \mathbb{I}\big[PPD_{i,c}(o) = DSL_{i,c}(o)\big].$$

To correct for agreement that can occur due to the marginal tendency of each source to disclose (or not disclose) categories, we compute the expected agreement (P_e) from the marginal disclosure rates:

$$p_1^{PPD}(o) = \frac{1}{N_o} \sum_{(i,c)\in\Omega_o} PPD_{i,c}(o), \quad p_1^{DSL}(o) = \frac{1}{N_o} \sum_{(i,c)\in\Omega_o} DSL_{i,c}(o),$$

with $p_0^{PPD}(o) = 1 - p_1^{PPD}(o)$ and $p_0^{DSL}(o) = 1 - p_1^{DSL}(o)$
The expected agreement is:

$$P_e(o) = p_1^{PPD}(o)\, p_1^{DSL}(o) + p_0^{PPD}(o)\, p_0^{DSL}(o)$$

Finally, Cohen's kappa is computed as:

$$\kappa(o) = \frac{P_o(o) - P_e(o)}{1 - P_e(o)} \tag{10}$$

We report κ score for collection and sharing to provide a chance-corrected measure of consistency between the two disclosure layers. We use κ in addition to percent agreement because both sources frequently report "not disclosed" across many categories, which can inflate raw agreement; κ discounts agreement attributable to these marginal base rates.

3.3.1 Sensitivity Risk Score (SRS). Building on the consistency indicators above, we define a privacy-sensitive risk score that weights each misalignment by the sensitivity of the underlying data category. This metric is crucial because data categories differ substantially in their potential for privacy harm. Accordingly, we adopt a standard weighted aggregation approach for combining heterogeneous dimensions with unequal importance. The resulting SRS represents the expected misalignment intensity under a sensitivity-weighted distribution over data categories. For instance, a mismatch involving a device model has a fundamentally different risk profile than a mismatch involving financial credentials or health status. Treating all discrepancies equally obscures the severity of disclosure failures. Additionally, a simple count of inconsistencies treats all violations equally, preventing the distinction between "inconvenient" and "dangerous". Therefore, weighting emphasizes the violations that matter most. Thus, our weighting misalignments metric, based on data sensitivity, produces a more realistic measure of disclosure risk than treating all inconsistencies equally.

Let C denote the set of data categories and let $w_c \in \{w_1, w_2, \ldots, w_n\}$ be a sensitivity weight for category $c \in C$. For an app i, category c, and operation $o \in \{\text{collect}, \text{share}\}$, we define the operation-specific SRS as:

$$\text{SRS}_i^{(o)} = \frac{\sum_{c\in C} w_c \big(\text{MisScore}_{i,c}^{(o)}\big)}{\sum_{c\in C} w_c}. \tag{11}$$

This resulting $\text{SRS}_i^{(o)}$ represents a normalized sensitivity-weighted average of disclosure misalignment intensity across data categories for an operation (sharing or collection). Categories with higher sensitivity weights contribute proportionally more to the overall score, ensuring that misalignment involving highly sensitive data e.g., financial or health information, has greater influence than misalignment involving low-risk data types.

To calculate the overall SRS for an app, we refer to $\text{SRS}_i^{(\text{share})}$ as SRS-S and $\text{SRS}_i^{(\text{collect})}$ as SRS-C. Thus, the overall scores is given as:

$$\text{SRS}_i^{(\text{overall})} = \frac{\text{SRS}_i^{(\text{share})} + \text{SRS}_i^{(\text{collect})}}{2} \tag{12}$$

Given that the SRS computes the expected misalignment intensity, it is important to emphasize sharing misalignment, as this implies that user data leave the device and therefore increases potential privacy risk. Accordingly, we define a revised overall score as follows:

$$\text{SRS}_i^{(\text{overall-w})} = \alpha\, \text{SRS}_i^{(\text{share})} + (1 - \alpha)\, \text{SRS}_i^{(\text{collect})} \tag{13}$$

where $0 \le \alpha \le 1$ is a threshold parameter that controls the relative importance of alignment.

4 Cross-Layer Privacy Analysis - Quantitative Findings

We now empirically evaluate the alignment between Google Play's Data Safety labels and app privacy policies relying on the methodology and dataset described in Section 3. This quantitative analysis is organized around the research questions presented in Section 2.2.

4.1 RQ1. How consistent are privacy policies and Google Play's Data Safety labels when reporting data collection and sharing?

We first quantify consistency between privacy policy document (PPD) and the Data Safety label (DSL) across apps, data categories, and data practices (collection vs. sharing). We assess consistency using (i) overall Consistency and Misalignment Scores for agreements and discrepancies, and (ii) the Cohen's kappa (κ) for chance-corrected agreement measure.

Overall Consistency and Misalignment: Using the Consistency Score defined in Equation 4, we computed the PPD and DSL agreement across all apps and data categories. Our results show that

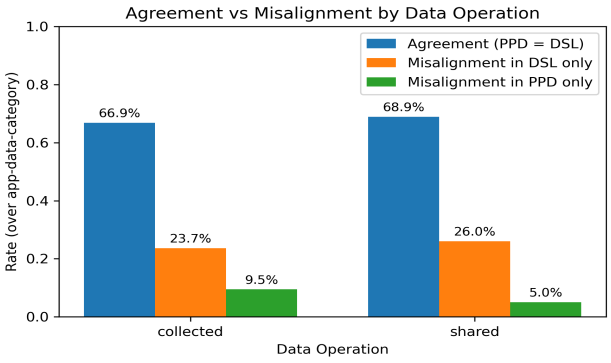

Figure 2: Overall agreement vs. misalignment rates between PPD and DSL by operation (collection vs. sharing), aggregated over all apps and data categories.

the PPD and DSL agree at about 66.9% for data collection disclosure and 68.9% for data sharing disclosures. Since misalignment as defined in Equation 8 is the complement of the consistency indicator as shown in Equation 9, correspondingly, the overall misalignment rates (prevalence of disagreement across all app–category pairs) for our dataset are 33.1% for data collection and 31.1% for data sharing. However, among misalignment, discrepancies are more often attributed to the *DSL only* than to the *PPD only*. For collection, 23.7% of disclosures are misaligned only in DSL versus 9.5% only in PPD; for sharing, the corresponding rates are 26.0% (DSL only) versus 5.0% (PPD only). This asymmetry indicates that DSL-only discrepancies are systematically more common, especially for data sharing. This micro-level misalignment prevalence is shown in Figure 2. At the macro-level, however, inconsistencies are more pervasive. Over 90.0% of apps contain at least one inconsistency in DSL-reported data collection, and 92.6% in DSL-reported data sharing. By comparison, inconsistencies in the PPD are substantially lower: 64.8% of apps show at least one collection inconsistency, and only 35.4% for sharing. This proportion reflects the significance of misalignment at the application level.

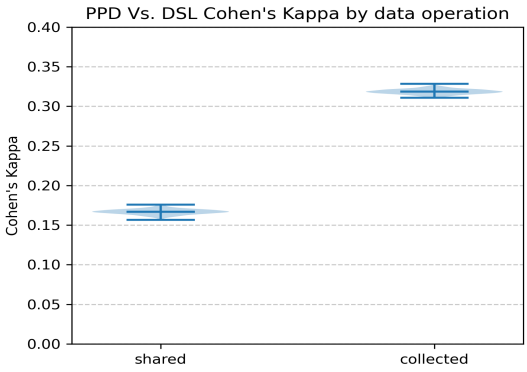

Figure 3: PPD vs DSL kappa score by data operation scope

Chance-corrected Consistency (Cohen's κ). Although the consistency measures computed above showed that about 2 out of 3 app-category pairs agree at the micro-level, this measure did not account for chance-based agreement. In our dataset, we found that both sources often report "no data collected" or "no data shared" for many of the 14 data categories; using the simple rate-based measure, this will result in high prevalence of negative answers and would otherwise inflate agreement even if the two disclosures are poorly aligned. The κ score, on the other hand, adjusts for this baseline and therefore provides a more reliable measure of true consistency across categories. Thus, to complement these rate-based prevalences, we compute Cohen's kappa (κ) to assess how consistently the PPD and DSL report collection and sharing, accounting for agreement that could occur by chance. Our result showed that the κ scores average around 31% for collection and 16% for sharing, as shown in Figure 3, indicating only weak agreement for collection and even weaker agreement for sharing. These low κ values suggest limited consistency between PPD and DSL, especially in sharing.

RQ1 Key Findings

Overall micro-level consistency rates are moderate (two-thirds), and chance-corrected consistencies are substantially lower (less than a third) for data sharing than for data collection. Additionally, macro-level inconsistencies are also widespread: almost all apps exhibit at least one disclosure misalignment. Taken together, these findings indicate that the two disclosure layers frequently present divergent representations of app data practices. Such divergence undermines the reliability of platform-level privacy disclosures and has significant implications for user data privacy, as users may base trust decisions on incomplete or inconsistent information.

4.2 RQ2. Which data categories are most frequently misaligned across the two disclosure layers?

While RQ1 examines overall micro and macro app-level consistency/misalignment between PPD and DSL at the collection and sharing, RQ2 focuses on identifying the specific data categories with the highest misalignment prevalence.

The heatmap in Figure 4 shows category-level total misalignment rates for collection and sharing. Several categories exhibit strikingly high rates. For collection, the worst categories are Photos and videos (51.6%), Location (51.1%), App activity (49.0%), App info and performance (44.0%), Financial info (43.9%), and Messages (41.1%). For sharing, misalignment is even more pervasive for some categories: Personal info (72.0%), Device or other IDs (55.1%), Location (49.6%), App info and performance (46.9%), App activity (41.5%), and Financial info (41.4%). In contrast, categories such as Audio, Calendar and Health and fitness have noticeably lower misalignment rates in both operations.

Additionally, to explore the specifics of the misalignment rate, we examine Misalignment in DSL (the Data Safety label omits collection/sharing described in the privacy policy) and Misalignment in PPD (the label claims collection/sharing that is not supported by the policy text) separately. Figure 5 focuses on Misalignment in DSL, and the result shows that there is a significant misalignment in highly sensitive data categories. For example, 69.6% of apps are misaligned in DSL for Personal info (shared). Similarly, for Location, misalignment in DSL is about 44.6% for collection and 42.6% for sharing; Device or other IDs (shared) show 48.8%, and Photos and

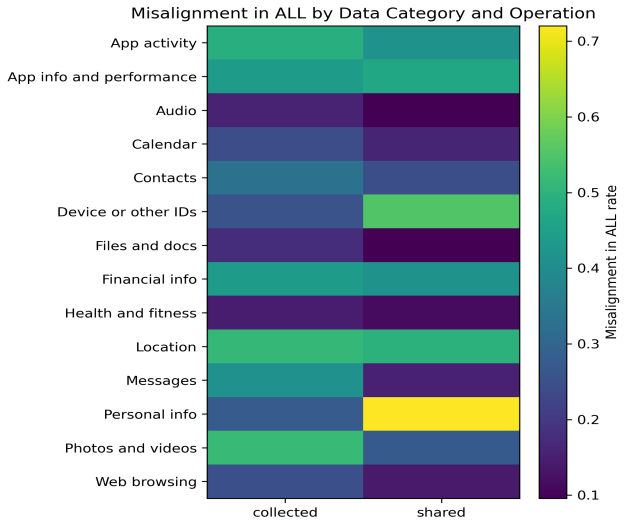

Figure 4: Misalignment rates by data category and operation scope.

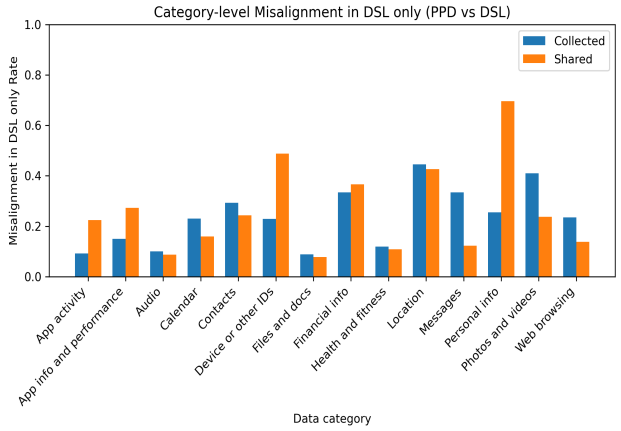

Figure 5: Misalignment rates by data category in Data Safety labels only.

videos (collected) have a 41.0% misalignment rate in DSL. Messages and Financial info have misalignment rates of about 33–37% in DSL, depending on the operation. Taken together, these patterns indicate that for sensitive categories, misalignment largely reflects omissions in the Data Safety label relative to the policy.

By contrast, Figure 6 shows that misalignment rate in the privacy policy document is much rarer and concentrated in a small set of categories. The most notable cases are App activity (collected), where 39.9% of apps are misaligned with privacy policies; App info and performance (collected) (29.0%); and Photos and videos (collected) and Financial info (collected), which both exceed 10%. For most other categories and for sharing in general, Misalignment in PPD stays below 7%.

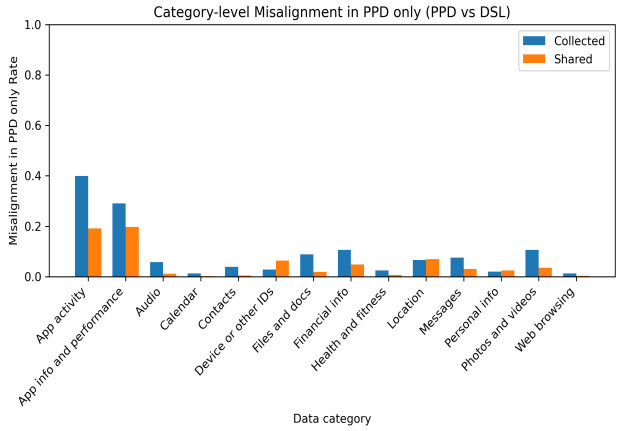

Figure 6: Misalignment rates by data category in privacy policies only.

Cosine Similarity. To complement the category-level misalignment analysis, we examine whether PPD and DSL disclosures exhibit similar *overall disclosure patterns* across the 14 data categories for both collection and sharing. We use cosine similarity to quantify *distributional alignment*, measuring how closely the disclosure profiles in the privacy policy and the Data Safety label align for a given app and data operation. Specifically, for each app and operation, we compute the cosine similarity between the vector of disclosed data-type terms (lexical) in the policy and the corresponding vector in the label. The resulting score ranges from 0 to 1, with higher values indicating greater overlap (agreement) in disclosed categories and lower values indicating misalignment. We summarize these similarity scores across apps using the empirical cumulative distribution function (ECDF), where each point on the curve represents the proportion of apps with similarity at or below that value.

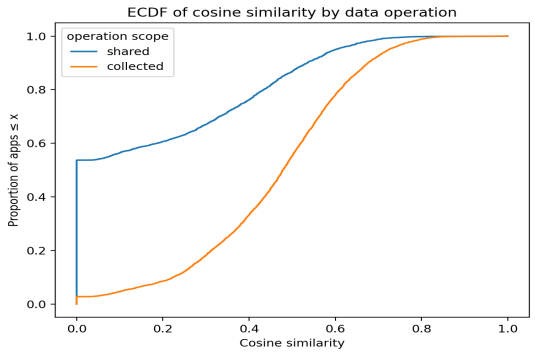

Figure 7: ECDF of cosine similarity between privacy policies and Data Safety labels, stratified by data operation scope.

As shown in Figure 7, the collection ECDF exhibits a sigmoidal (S-shaped) pattern, indicating that cosine similarity scores are concentrated around a central range rather than being uniformly distributed or strongly skewed. This suggests that, for most apps, Data Safety collection disclosures exhibit moderate structural alignment

with the corresponding privacy policy descriptions. By contrast, the sharing curve rises sharply at very low similarity values, with approximately half of apps exhibiting cosine similarity near zero. This indicates a substantial subset of apps where reported sharing behavior in the Data Safety label has little to no lexical overlap with the privacy policy. This pattern is consistent with the frequent use of "No data shared with third parties" declarations in the Data Safety section, even when policies describe integrations with external partners. The median cosine similarity for collection is approximately 0.55, whereas the sharing distribution is heavily concentrated near the lower bound. Overall, developers appear substantially more consistent in describing data collection practices than data sharing practices. These aggregate trends are further illustrated by the case study in Appendix F.1, where a "No data shared with third parties" coexists with extensive data sharing with external partners.

RQ2 Key Findings

Overall, we find that high-sensitivity data categories exhibit the greatest misalignment between privacy policies and Data Safety labels. These discrepancies are driven primarily by under-reporting in the Data Safety label rather than over-disclosure relative to the policy. Consistent with this result, cosine similarity analysis also shows stronger distributional alignment for data collection than for data sharing, indicating that developers describe what apps collect more consistently than what they share. Because the most sensitive data categories are disproportionately affected—and sharing disclosures are particularly inconsistent—users may receive incomplete or misleading information about how and with whom their data are shared, undermining the reliability of data disclosures.

4.3 RQ3. How are apps distributed across low, medium, and high Sensitivity Risk Score tiers, and what does this distribution reveal about the severity of privacy misalignment between privacy policies and Data Safety labels?

RQ1 and RQ2 establish the prevalence of disclosure inconsistencies at both the app and data-category levels. We now extend this analysis by evaluating the privacy-sensitive impact of misalignment using the Sensitivity Risk Score (SRS) as defined in Equation 11, which weights discrepancies according to the potential harm of the underlying data categories.

We assign sensitivity weights $w_c \in \{1, 2, 3\}$ corresponding to low, medium, and high sensitivity, to each of the 14 Google Play data categories. The categorization into three tiers is informed by prior literature characterizing the relative privacy risk associated with these data types. First, we treated *Personal info* as highly sensitive, since prior disclosure-scoring work [2] assigns higher weights to core identifiers such as name, address, and contact details than to generic profile attributes. Next, we leverage the work of Sardana et al. [46] and Chang et al. [13], which assign much higher sensitivity scores to permissions that access location, contact, financial, and health data than to diagnostics or performance metrics, to inform additional grouping decisions. Finally, drawing on the work of Gómez Ortega et al. [19], which shows that voice interaction histories are often perceived as sensitive, we assigned audio a high sensitivity score in our scoring. Table 3 summarizes the final mapping used in our Sensitivity Risk Score, following the prior work characterizations.

Table 3: Sensitivity weights assigned to Google Play data categories.

Weight	Data categories
3	Location, Personal info, Financial info, Audio, Contacts, Health and fitness, Messages, Photos and videos, Web browsing, Device or other IDs
2	Files and docs, App activity
1	Calendar, App info and performance

For the sensitivity risk scoring, we interpret each $\text{SRS}_i^{(o)}$ variants: *SRS−C* (collection) and *SRS−S* (sharing), and *SRS−O−w* (the overall weighted score -$\text{SRS}_i^{(\text{overall-w})}$) on a 0–1 scale. To facilitate interpretation, we categorize apps into three risk tiers: low (score < 0.30), medium ($0.30 \leq$ score < 0.70), and high (score ≥ 0.70). For the overall score defined in Equation 13, we set the parameter α, which assigns moderately greater weight to sharing disclosures. As discussed in Section 3.3.1, this weighting reflects the greater privacy risk associated with transmitting data to third parties compared to collection alone. To evaluate the robustness of this choice, we conducted a sensitivity analysis using multiple α values, as reported in Appendix H. The results show that the overall risk stratification remains stable across the evaluated settings, with $\alpha = 0.6$ providing a balanced emphasis between collection and sharing disclosures. Using $\alpha = 0.6$, we computed the final SRS values for our dataset. A descriptive example of the SRS computation is provided in Appendix D, and the resulting score distribution is illustrated in Figure 8.

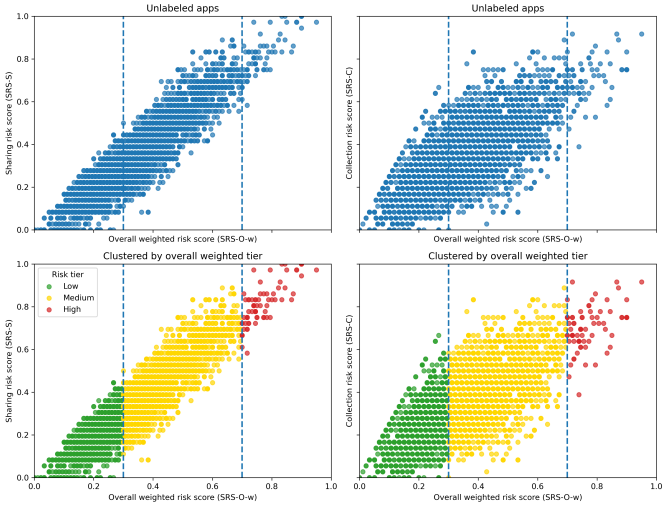

Figure 8: Distribution of apps across risk tiers.

The visualization in Figure 8 shows the distribution of Sensitivity Risk Scores across all 6,051 apps by plotting the overall risk score ($SRS - O - w$) on the x-axis and the data collection and data sharing

scores ($SRS - C$ and $SRS - S$) on the y-axes. Under $SRS - O - w$, apps are distributed between the low- and medium-risk tiers, with 49.61% of apps in the low-risk tier and 49.10% in the medium-risk tier, while only 1.29% fall in the high-risk tier. Accordingly, the distribution is concentrated in the low- and medium-risk regions, with a comparatively small but visible tail extending into the high-risk region. Apps located in the high-risk region generally exhibit elevated data sharing risk ($SRS - S$), and many also exhibit elevated data collection risk ($SRS - C$). This pattern suggests that high overall sensitivity-weighted risk typically reflects concentrated risk across multiple sensitive data practices, rather than being attributable to a single isolated factor. When comparing risk by data practice, high data-sharing risk is more prevalent than high data-collection risk. Specifically, 177 apps (2.93%) fall in the high-risk tier under $SRS - S$, whereas 86 apps (1.42%) fall in the high-risk tier under $SRS - C$. Comparing RQ3 with RQ2—which shows that high-sensitivity data categories are disproportionately affected by disclosure misalignment—the sensitivity-weighted SRS analysis reveals that substantial cumulative privacy risk is concentrated in a smaller subset of apps. In other words, while high-risk categories are more prone to misalignment, only a limited fraction of applications exhibit sustained or multi-category discrepancies severe enough to elevate their overall risk tier.

Importantly, the high-risk tier reflects not only the *extent* of misalignment but also its *sensitivity*. Under the weighting scheme in Table 3, apps enter the high-risk region when discrepancies involve data categories with greater privacy harm potential. From a user-impact perspective, elevated risk in high-sensitivity categories—such as location, financial, health, communication, and identifier-related data—can increase exposure to profiling, identity linkage, behavioral tracking, discrimination, financial fraud, and disclosure of intimate personal attributes. Accordingly, the high-risk tier serves as a practical prioritization signal for auditing, developer review, and platform-level enforcement.

RQ3 Key Findings

Overall, our results show that sensitivity-weighted scoring yields a stable and interpretable stratification of privacy risk across apps, while distinguishing how risk is distributed between data sharing and data collection. Notably, the distribution exhibits a more pronounced high-risk tail for data sharing than for data collection. These findings support $SRS_i^{(\text{overall-w})}$ as an effective measure of app-level privacy risk and demonstrate the value of tier-based stratification for prioritizing review of apps with risk concentrated in highly sensitive data categories.

4.4 RQ4. Which app categories exhibit the highest misalignment risk, and how is risk distributed within those categories?

To investigate whether sensitivity risks are concentrated in particular kinds of apps, we analyze the SRS across Google Play app categories. For each category, we compute the mean overall weighted score $SRS_i^{(\text{overall-w})}$ and the distribution of apps across the low, medium, and high risk tiers. Figure 9 shows the top 20 categories ranked by mean overall weighted SRS. All of these categories have average scores between roughly 0.43 and 0.56, indicating that a typical app in these segments exhibits nontrivial misalignment between its privacy policy and data safety label. Categories related to communication and continuous data collection—such as *wearables*, *communication*, and *ANDROID_WEAR*—appear near the top of the ranking, as do several health- and lifestyle-related categories (*sleep tracker*, *fitness*, *wellness*). We also observe elevated scores for utility categories that can access devices or accounts remotely (e.g., *remote desktop*, *EV charging locator*), suggesting that apps with privileged or always-on access to users and devices are more prone to inconsistent disclosure.

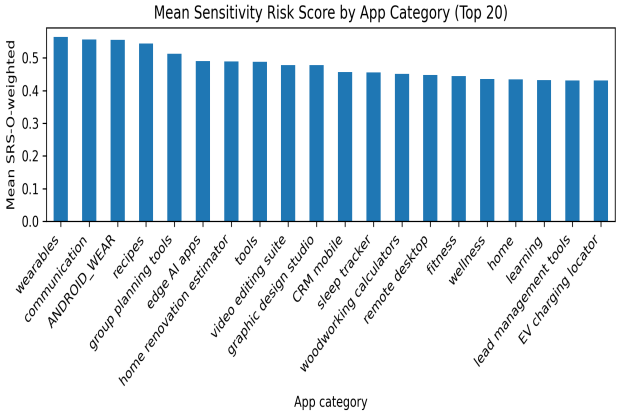

Figure 9: Risk scores across app categories.

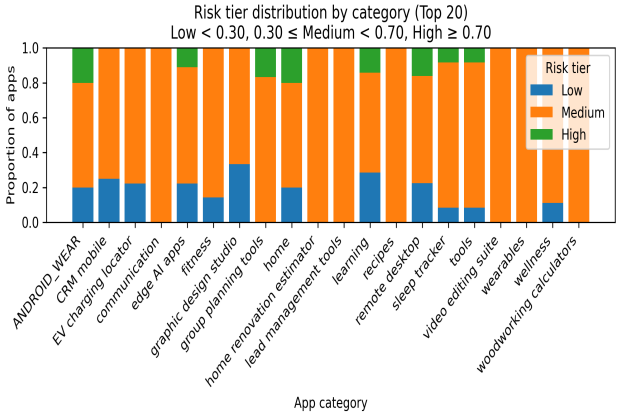

Figure 10: Distribution of app categories across risk tiers.

Figure 10 represents how apps are distributed across risk tiers within the top app categories shown in Figure 9. In almost every category, the medium-risk tier dominates, confirming that moderate PPD–DSL misalignment is not an isolated occurrence. However, the share of high-risk apps is not uniform: categories such as *ANDROID_WEAR*, *EV charging locator*, *group planning tools*, and *home renovation estimator* contain a visibly larger fraction of high-risk apps compared to others, while categories like *wellness* or

woodworking calculators are almost entirely composed of low- and medium-risk apps.

RQ4 Key Findings

Using the sensitivity-weighted privacy risk score, we observe that app categories capable of persistent user monitoring, communication mediation, or device control exhibit significantly higher risk. This suggests that disclosure misalignment within these app types disproportionately involves high-sensitivity data categories, highlighting their potential privacy impact and underscoring the need for careful scrutiny of privacy disclosure claims in these domains.

Additional quantitative evaluation of the correlation between privacy misalignment risk and app popularity is provided in Appendix E.

5 Qualitative Findings

5.1 Shared but Not Collected: Data Safety Labels

In our subset of 3,111 apps without a privacy policy URL, we observed a recurring anomaly in which Google Play's Data Safety label reports that an app shares user data while simultaneously declaring that it does not collect any data. Among these apps, 24.36% indicate that user data are "shared" but that "no data" are collected. An example is *GlobeOne: Get More from Globe*, whose store listing states that the app "may share" location, personal information, and device or other IDs with third parties, but also shows the summary "No data collected" [17]. According to Google's own documentation, the Data Safety section is meant to disclose how apps collect, share, and protect user data, and developers are required to complete a form describing their practices, which is then shown on the store listing. Google explains that the label is developer-provided, and that "data collected" covers user data transmitted off the device, while "data shared" refers to data that is passed on to third parties [20], [8]. Under these definitions, a pattern where multiple categories of personal and device data are marked as "shared" but overall "no data collected" appears logically inconsistent and suggests either misunderstanding of the form or misreporting of practices by the developer [48]. This type of labeling anomaly is important for our analysis because it illustrates how developer-provided disclosures can undermine the intended transparency of the Data Safety section and obscure the true extent of applications data-handling practices.

5.2 The Service-Provider Loophole in Data Sharing Disclosures

Google Play's Data Safety documentation allows developers to treat certain third parties as "service providers" rather than as recipients of "shared" data, as long as those entities process data solely on the developer's behalf [20]. For example, an analytics platform or cloud hosting provider can be classified as a service provider, and the associated data flows do not need to be disclosed as "sharing" in the Data Safety form. By contrast, Google explicitly states that if an SDK aggregates data across multiple customers to build advertising profiles, this is not service-provider activity and must be declared as data sharing in the label. In practice, this distinction can produce labels that say "No data shared with third parties" even when data is sent to large external companies. MX Player provides a concrete example: its privacy policy describes integrations with analytics, advertising, and SDK partners such as Google Analytics, Facebook Ads/AdMob, YouTube API Services, and Google APIs that collect information about installed apps and support interest-based advertising and profiling. If the developer classifies all of these partners as service providers, the Play Store listing can still show "No data shared with third parties" in the Data Safety summary, even though user data flows to those platforms [6, 35].

Under Google's own guidance, however, any cross-app profiling or reuse of data across customers should be treated as sharing, implying that a "No data shared" label would be misleading in such a scenario. This pattern is consistent with our large-scale analysis in Section 4, where we find that sharing disclosures are much less aligned with privacy policies than collection disclosures. This highlights a gap in how Google supports both users and developers. For users, the current label design does not reveal which external companies receive their data or whether those companies reuse it across apps, making it difficult to understand the true extent of data disclosure. For developers, the burden of correctly interpreting nuanced concepts such as "service provider" and "cross-app profiling" is placed entirely on them, while Google simultaneously emphasizes that it does not systematically verify the accuracy of declarations [20]. We also observed data types not captured by the Data Safety categories; examples are provided in Appendix I.

5.3 Recommendations

Our findings underscore the importance of strengthening Data Safety disclosure mechanisms on platforms such as Google Play. Ensuring closer alignment with privacy policies—particularly for sensitive data categories—is critical to reducing privacy-sensitive disclosure risk and providing users with an accurate understanding of data practices.

5.3.1 Implement Automated Consistency and Sensitivity Scoring at Submission Time. Beyond strengthening disclosure design, we recommend that Google integrate an automated cross-layer consistency verification mechanism at the time of app submission. Specifically, the platform could: (i) Automatically compare a developer's privacy policy against the submitted Data Safety label using a structured category mapping similar to the methodology presented in this study. (ii) Compute and internally validate a Consistency Score and a Sensitivity Risk Score (SRS) that quantify the degree and risk-weighted impact of disclosure misalignment. (iii) Surface these scores transparently to users as part of the app listing interface. Displaying a consistency indicator would provide users with an interpretable signal of disclosure reliability. Over time, greater visibility of low-consistency or high-risk scores may incentivize developers to improve alignment between privacy policies and Data Safety labels, creating market-driven pressure toward more accurate and trustworthy disclosures.

5.3.2 Strengthen the Design and Transparency of the Data Safety Framework. The results from this study indicate that disclosure misalignment—particularly involving sensitive data categories—often stems from limited transparency around third-party SDK practices and ambiguous service-provider classifications. To address this, we

recommend that Google strengthen the design and enforcement of the Data Safety framework through the following measures: (i) Require explicit identification of analytics, advertising, and other high-impact SDK partners within the Data Safety form, even when developers classify them as service providers. This should include disclosure of whether such partners reuse data across multiple applications. (ii) Surface SDK-level data accesses directly within the Data Safety interface, linking specific data categories (e.g., location, identifiers, financial data) to the third parties that receive them. This prevents developers from generically labeling all partners as service providers without concise differentiation. (iii) Establish standardized incident reporting mechanisms that map exposed data fields back to the exact categories and collection/sharing operations presented in the Data Safety label, improving accountability and post-incident transparency.

5.4 Limitations and Future Work

While our framework enables large-scale analysis of cross-layer privacy disclosure consistency, several limitations should be considered when interpreting the results. In this context, terms such as "under-reporting" refer to relative disclosure differences between the two artifacts; for example, when a data category is disclosed in the privacy policy but omitted from the corresponding Data Safety label. Such discrepancies do not necessarily establish that the label is factually incorrect or that the application actually collects or shares the data in practice. Rather, they indicate divergence between two developer-provided representations of data practices. Our privacy-policy analysis relies on an LLM-based extraction pipeline. We validated the pipeline on a manually labeled subset and observed stable performance across two high-performing LLM backends as shown in Section 3.1.3. Nevertheless, extraction quality may vary across models, policy-writing styles, ambiguous legal language, or future model versions. In addition, our crawler is currently restricted to English-language privacy policies, which may limit generalizability to non-English, multilingual, or region-specific app ecosystems where disclosures may be expressed differently. However, because our framework relies on general-purpose LLM extraction, extending the pipeline to multilingual settings is a feasible direction for future work. Finally, our framework intentionally represents disclosures using category-level binary indicators over the 14 Google Play data categories to support consistent, scalable, and reproducible comparison across thousands of apps. While appropriate for measuring broad cross-layer disclosure alignment, this abstraction does not capture all contextual nuances present in privacy disclosures, such as processing purpose, optional versus required collection, retention conditions, user controls, or fine-grained distinctions within broader data categories. Consequently, some partial or conditional disclosures may not be fully represented in the binary encoding.

As future work, we plan to extend our framework to a three-way analysis that triangulates static code analysis, privacy policies, and Data Safety labels. Such an approach could help distinguish benign disclosure inconsistencies from systematic under-disclosure and support more targeted privacy auditing and platform enforcement. Another important direction is longitudinal measurement. By periodically re-crawling the same applications, future work could examine how Data Safety labels and privacy policies evolve over time in response to policy changes, enforcement actions, platform redesigns, or public privacy incidents. Finally, future work could extend this cross-layer analysis to the Apple ecosystem by comparing Apple Privacy Nutrition Labels with privacy policies and, where possible, analyzing cross-platform versions of the same applications.

6 Related Work

In this section, we review the related literature in automated privacy policy analysis and measurement studies evaluating the accuracy and usability of app-store privacy labels.

6.1 Privacy Policy Analysis

Early efforts to automate privacy policy analysis used NLP and conventional ML to reduce the burden of reading long policies and to support scalable auditing. The development of annotated corpora proved fundamental to this research direction, with Wilson et al. [53] introduced the OPP-115 dataset of 115 policies labeled by law students, which became a widely used benchmark for training and evaluating policy understanding systems [54]. Building on such resources, several systems extracted structured representations of data practices. PolicyLint [7] used rule-based NLP to detect inconsistencies and potential contradictions in policy statements, while PrivOnto [37] mapped expert annotations into an ontology to support query knowledge bases of collection, sharing, and purpose. Yu et al. [57] similarly leveraged syntactic parsing and pattern matching to surface critical sentences about data handling. In parallel, statistical classifiers framed policy analysis as supervised text classification or completeness assessment. Costante et al. [14] and Guntamukkala et al. [21] applied multiple ML models to categorize policy content and assess coverage relative to regulatory-inspired categories. Later systems improved scalability and output structure such as Polisis [22] used hierarchical neural models to map policy text into structured, human-readable labels, and PrivacyCheck [58] used TF–IDF features with gradient-boosted classifiers to generate aggregate privacy scores. Other work targeted specific policy elements, such as opt-out and choice language [47]. In general, these prior methods rely on task-specific schemas, extensive manual annotation, and brittle rule-based extraction pipelines.

These limitations have motivated a new line of research leveraging LLM-assisted policy analysis frameworks to automatically extract data collection and sharing statements from privacy policies into normalized schemas suitable for large-scale measurement. The works of [18, 41, 42, 49, 56, 58] have shown strong LLM performance in policy summarization. Additionally, [5, 10, 12, 16, 34, 51] proposed the use of LLMs for regulatory-focused policy assessment and compliance checking from privacy policy. Woodring et al.'s *Privacify* [55] leverages LLMs to summarize policies and support regulation-oriented analysis (e.g., GDPR, HIPAA, COPPA, FERPA) by generating structured outputs and legal rationales. Rodriguez et al. [45] further demonstrate that general-purpose LLMs such as ChatGPT and Llama 2 can compete with and often outperform state-of-the-art traditional methods for extracting privacy practice disclosures, suggesting a paradigm shift towards efficient and accessible analysis. Beyond extraction and regulatory mapping, LLMs

can also support user-centered understanding by mitigating information overload; for example, Mori et al. [33] use LLMs to evaluate and improve policy understandability, showing that these models can identify policy descriptions that users are likely to misunderstand. Tang et al. [50] further introduced *PolicyGPT*, an LLM-based framework for classifying policies with respect to GDPR requirements. Our work builds on this foundation of LLM-assisted policy analysis to investigate cross-layer consistency between privacy policies and platform-level Data Safety disclosures.

In contrast, we move beyond structured extraction by extending the LLM-assisted policy analysis framework proposed by [55] to support systematic cross-layer disclosure alignment measurement. In addition, we introduce a sensitivity-weighted scoring framework that quantifies the privacy risk associated with disclosure misalignment.

6.2 Data Safety Labels Evaluation

With the mandatory introduction of short-form privacy disclosures, including Apple's Privacy Nutrition Labels and Google Play's Data Safety section, recent research has examined whether these labels serve as reliable transparency mechanisms by evaluating their accuracy and how effectively they support users' privacy decisions. On iOS, Ali et al. [3] used BERT-based language models to compare Apple privacy labels against privacy policies and found widespread mismatches in which policies often described broader data collection than what labels reported. Using behavioral evidence instead of policies, Kollnig et al. [29] evaluated Privacy Nutrition Labels using static analysis and network-traffic monitoring, showing that apps labeled as not collecting data can still include tracker libraries and exhibit tracking-related behavior. Cross-ecosystem comparisons further reveal inconsistency even for the same app: Rodriguez et al. [44] compared disclosed data collection across five data categories and found that iOS and Google Play labels frequently disagree. Focusing specifically on Android, Khandelwal et al. [25] conducted the first comprehensive measurement of Google Play's Data Safety Section, identifying substantial internal inconsistencies and evidence of both under- and over-reporting. To explain how such issues arise, their developer study highlights practical reporting challenges such as ambiguous terminology, difficulty accounting for SDK-driven data practices, and the burden of keeping disclosures updated, which aligns with Li et al. [31], who observed and interviewed developers completing Apple's labels and found that misunderstandings of label terms and uncertainty about what to disclose are common sources of inaccuracies. Beyond correctness, user studies suggest that labels often fail to meet user information needs. Zhang et al. [61] showed that current labels answer only a minority of users' real privacy questions, particularly omitting details such as who receives shared data and who can access it, while Zhang et al. [59] found that users struggle with label structure and technical terminology and that labels have limited impact on users' ability to manage privacy. In contrast to prior work, our study introduces a cross-layer privacy disclosure analysis pipeline that normalizes privacy-policy and Data Safety disclosures into a unified schema, enables large-scale alignment analysis of data collection and sharing practices, and ranks disclosure misalignment using a sensitivity-weighted risk score.

6.3 Cross-Layer Privacy Disclosure Studies

Recent studies, including Alomar et al. [4] and Jain et al. [23], have examined inconsistencies between platform privacy labels and privacy policies in both Android and iOS ecosystems. ATLAS [23] studies discrepancies between Apple Privacy Labels and privacy policies at scale in the iOS ecosystem using an ensemble classifier for privacy-label prediction, while Alomar et al. [4] analyze behavioral compliance and inaccurate disclosures in child-directed Android apps using dynamic traffic analysis and SDK inspection. Our work differs in both analytical scope and methodological focus. We focus on cross-layer consistency between privacy policies and Google Play Data Safety Labels across 6,051 Android apps using a unified 14-category Google Play schema. Methodologically, we leverage an LLM-based extraction framework to identify explicit collection and sharing disclosures directly from privacy policies and compare them against developer-declared DSL disclosures.

Beyond identifying disclosure discrepancies, our study introduces additional analytical dimensions for characterizing disclosure misalignment. First, we explicitly distinguish data collection and data sharing as separate disclosure operations throughout the analysis. This enables operation-specific consistency measurements and reveals a strong asymmetry between collection and sharing disclosures, with sharing disclosures exhibiting substantially lower consistency. Second, our framework decomposes discrepancies into DSL-only and PPD-only mismatches, enabling characterization of systematic omission patterns within Google Play Data Safety Labels, particularly for sharing-related disclosures. Third, we introduce a Sensitivity Risk Score that weights discrepancies according to the sensitivity of the underlying data categories, allowing misalignment severity to be analyzed beyond raw discrepancy frequency alone. Finally, we incorporate chance-corrected agreement analysis using Cohen's (κ) and structural disclosure alignment using cosine similarity to characterize agreement beyond raw match rates.

Collectively, these analyses provide a more fine-grained and risk-oriented characterization of cross-layer disclosure misalignment across the Google Play ecosystem, revealing that sharing disclosures are systematically less aligned than collection disclosures and that elevated disclosure risk disproportionately concentrates in high-sensitivity data categories.

7 Conclusion

This study examined how Android developers' privacy policies align with their Google Play Data Safety labels regarding user data-handling disclosures. Using a unified schema across 14 data categories and a sensitive privacy risk score, our analysis of over 6000 apps reveals widespread inconsistencies between privacy policies and data safety labels, particularly regarding data sharing. Under-reporting in the data safety label is much more common, and mismatches are concentrated in sensitive categories such as personal information, location, financial data, device identifiers, and messages. Our risk analysis further indicates that around half of the apps fall into a medium-risk tier, and a fraction, but an important subset, appear high risk. Overall, current developer self-reported labels provide only a partial and sometimes misleading picture of user data practices, underscoring the need for stronger platform checks and more transparent, verifiable disclosure mechanisms.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors used generative AI-based tools to revise the text, correct typographical errors, and grammatical errors.

References

[1] AccuWeather, Inc. 2025. AccuWeather Privacy Statement. https://www.accuweather.com/en/privacy. Accessed: 27 October. 2025.

[2] Erfan Aghasian, Saurabh Garg, Longxiang Gao, Shui Yu, and James Montgomery. 2017. Scoring users' privacy disclosure across multiple online social networks. *IEEE access* 5 (2017), 13118–13130.

[3] Mir Masood Ali, David G Balash, Monica Kodwani, Chris Kanich, and Adam J Aviv. 2024. Honesty is the best policy: On the accuracy of apple privacy labels compared to apps' privacy policies. In *Proceedings on Privacy Enhancing Technologies*, Vol. 2024. 142–166.

[4] Noura Alomar, Joel Reardon, Aniketh Girish, Narseo Vallina-Rodriguez, Serge Egelman, et al. 2025. The effect of platform policies on app privacy compliance: A study of child-directed apps. *Proceedings on Privacy Enhancing Technologies 2025* 3 (2025).

[5] Orlando Amaral, Sallam Abualhaija, Damiano Torre, Mehrdad Sabetzadeh, and Lionel C Briand. 2021. AI-enabled automation for completeness checking of privacy policies. *IEEE Transactions on Software Engineering* 48, 11 (2021), 4647–4674.

[6] Amazon Mobile LLC. 2025. MX Player — Data safety. https://play.google.com/store/apps/datasafety?id=com.mxtech.videoplayer.ad&hl=en_US. Accessed: 28 October 2025.

[7] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. {PolicyLint}: investigating internal privacy policy contradictions on google play. In *28th USENIX security symposium (USENIX security 19)*. 585–602.

[8] Android Developers. 2022. *Declare your app's data use.* https://developer.android.com/privacy-and-security/declare-data-use Accessed: 27 Nov. 2025.

[9] Android Developers. 2025. Terms of Service and data safety — Play Integrity. https://developer.android.com/google/play/integrity/terms. Accessed: 30 Oct. 2025.

[10] Irem Aydin, Hermann Diebel-Fischer, Vincent Freiberger, Julia Möller-Klapperich, Erik Buchmann, Michael Färber, Anne Lauber-Rönsberg, and Birte Platow. 2024. Assessing privacy policies with AI: ethical, legal, and technical challenges. *arXiv preprint arXiv:2410.08381* (2024).

[11] California Legislature. 2018. California Consumer Privacy Act of 2018. California Civil Code § 1798.100 et seq.. https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=1798.100 Also known as the California Consumer Privacy Act (CCPA), as amended, including by the California Privacy Rights Act of 2020 (CPRA).

[12] Rupert Chadwick, Sophie Blundell, and Emily Prendergast. 2024. Assessments of the Privacy Compliance in Commercial Large Language Models. https://doi.org/10.31219/osf.io/ugmfw Preprint. Accessed: 27 Nov. 2025.

[13] Kai Chih Chang, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2020. A framework for estimating privacy risk scores of mobile apps. In *International Conference on Information Security*. Springer, 217–233.

[14] Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry den Hartog. 2012. A machine learning solution to assess privacy policy completeness. In *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*. 91–96.

[15] European Parliament and of the Council. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88. https://gdpr-info.eu/ Informal consolidation of the General Data Protection Regulation (GDPR) for quick reference..

[16] Vincent Freiberger, Arthur Fleig, and Erik Buchmann. 2025. " You don't need a university degree to comprehend data protection this way": LLM-Powered Interactive Privacy Policy Assessment. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[17] Globe Telecom. 2025. *GlobeOne: Get More from Globe – Apps on Google Play.* https://play.google.com/store/apps/datasafety?id=ph.com.globe.globeonesuperapp Data Safety section.

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

[19] Alejandra Gómez Ortega, Jacky Bourgeois, and Gerd Kortuem. 2023. What is sensitive about (sensitive) data? Characterizing sensitivity and intimacy with Google assistant users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[20] Google. 2025. *Provide information for Google Play's Data safety section.* https://support.google.com/googleplay/android-developer/answer/10787469 Accessed October 29, 2025.

[21] Niharika Guntamukkala, Rozita Dara, and Gurdip Grewal. 2015. A machine-learning based approach for measuring the completeness of online privacy policies. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 289–294.

[22] 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 *27th USENIX Security Symposium (USENIX Security 18)*. 531–548.

[23] Akshath Jain, David Rodriguez, Jose M Del Alamo, and Norman Sadeh. 2023. Atlas: Automatically detecting discrepancies between privacy policies and privacy labels. In *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 94–107.

[24] Rishabh Khandelwal, Asmit Nayak, Paul Chung, and Kassem Fawaz. 2023. Comparing privacy labels of applications in android and ios. In *Proceedings of the 22nd Workshop on Privacy in the Electronic Society*. 61–73.

[25] Rishabh Khandelwal, Asmit Nayak, Paul Chung, and Kassem Fawaz. 2024. Unpacking privacy labels: A measurement and developer perspective on google's data safety section. In *33rd USENIX Security Symposium (USENIX Security 24)*. 2831–2848.

[26] Mst Eshita Khatun, Lamine Noureddine, Sideeq Bello, and Aisha Ali-Gombe. 2026. Artifacts_PPD_DSL: Artifact for "Disclosure Divergence: Measuring Privacy Policy and Data Safety Misalignment at Scale". https://github.com/Eshita66/Artifacts_PPD_DSL. GitHub repository, accessed June 2026.

[27] Mugdha Khedkar, Ambuj Kumar Mondal, and Eric Bodden. 2024. Do Android app developers accurately report collection of privacy-related data?. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering Workshops*. 176–186.

[28] Simon Koch, Malte Wessels, Benjamin Altpeter, Madita Olvermann, and Martin Johns. 2022. Keeping privacy labels honest. *Proceedings on Privacy Enhancing Technologies* (2022).

[29] Konrad Kollnig, Anastasia Shuba, Max Van Kleek, Reuben Binns, and Nigel Shadbolt. 2022. Goodbye Tracking? Impact of iOS App Tracking Transparency and Privacy Labels. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 508–520. https://doi.org/10.1145/3531146.3533116

[30] Renuka Kumar, Apurva Virkud, Ram Sundara Raman, Atul Prakash, and Roya Ensafi. 2022. A large-scale investigation into geodifferences in mobile apps. In *31st USENIX Security Symposium (USENIX Security 22)*. 1203–1220.

[31] Tianshi Li, Kayla Reiman, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I Hong. 2022. Understanding challenges for developers to create accurate privacy nutrition labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[32] Shuang Liu, Baiyang Zhao, Renjie Guo, Guozhu Meng, Fan Zhang, and Meishan Zhang. 2021. Have you been properly notified? automatic compliance analysis of privacy policy text with gdpr article 13. In *Proceedings of the Web Conference 2021*. 2154–2164.

[33] Keika Mori, Daiki Ito, Takumi Fukunaga, Takuya Watanabe, Yuta Takata, Masaki Kamizono, and Tatsuya Mori. 2025. Evaluating LLMs Towards Automated Assessment of Privacy Policy Understandability. In *Proceedings of the 2025 Symposium on Usable Security and Privacy*.

[34] Keika Mori, Tatsuya Nagai, Yuta Takata, and Masaki Kamizono. 2022. Analysis of privacy compliance by classifying multiple policies on the web. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 1734–1741.

[35] MX Media and Entertainment Pte Ltd. n.d.. MX Player Privacy Policy. https://mx.j2inter.com/about/privacy-policy. Accessed: 28 October 2025.

[36] Razieh Nokhbeh Zaeem, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava, and K Suzanne Barber. 2020. PrivacyCheck v2: A tool that recaps privacy policies for you. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 3441–3444.

[37] Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B Norton, N Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. 2018. PrivOnto: a semantic framework for the analysis of privacy policies. *Semantic Web* 9, 2 (2018), 185–203.

[38] OpenAI. 2025. ChatGPT — Data safety. https://play.google.com/store/apps/datasafety?id=com.openai.chatgpt&hl=en_US. Accessed: 28 November 2025.

[39] OpenAI. 2025. Privacy Policy. https://openai.com/policies/row-privacy-policy/. Accessed: 28 November 2025.

[40] OpenAI. 2025. What to Know About a Recent Mixpanel Security Incident. https://openai.com/index/mixpanel-incident/. Accessed: 28 November 2025.

[41] Przemysław Pałka, Francesca Lagioia, Rūta Liepina, Marco Lippi, and Giovanni Sartor. 2025. Make privacy policies longer and appoint LLM readers. *Artificial Intelligence and Law* (2025), 1–33.

[42] Jielin Qiu, Mengdi Xu, William Han, Seungwhan Moon, and Ding Zhao. 2024. Embodied executable policy learning with language-based scene summarization.

In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1896–1913.

[43] David Rodriguez, Joseph A Calandrino, Jose M Del Alamo, and Norman Sadeh. 2025. Privacy Settings of Third-Party Libraries in Android Apps: A Study of Facebook SDKs. *Proceedings on Privacy Enhancing Technologies* (2025).

[44] David Rodriguez, Akshath Jain, Jose M Del Alamo, and Norman Sadeh. 2023. Comparing privacy label disclosures of apps published in both the App Store and Google Play Stores. In *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 150–157.

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

[46] Neetu Sardana and Arpita Jadhav Bhatt. 2023. iPDS: Computing Privacy Disclosure Score for iOS apps and detection of privacy-infringing apps using Machine learning classifiers. In *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*. 699–705.

[47] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2774–2779.

[48] Harsh Shah. 2022. Google Play Data Safety: What You Need To Know. https://clevertap.com/blog/google-play-data-safety/. Accessed: 15 Nov. 2025.

[49] Sulayman Sowe, Tobias Kiel, Alexander Neumann, Yongli Mou, Vassilios Peristeras, and Stefan Decker. 2024. The Design and Implementation of APLOS: An Automated PoLicy DOcument Summarisation System. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. IEEE, 345–356.

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

[51] Damiano Torre, Sallam Abualhaija, Mehrdad Sabetzadeh, Lionel Briand, Katrien Baetens, Peter Goes, and Sylvie Forastier. 2020. An ai-assisted approach for checking the completeness of privacy policies against gdpr. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*. IEEE, 136–146.

[52] Liu Wang, Dong Wang, Shidong Pan, Zheng Jiang, Haoyu Wang, and Yi Wang. 2025. A big step forward? a user-centric examination of ios app privacy report and enhancements. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 4210–4228.

[53] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 1330–1340.

[54] Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, et al. 2018. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Transactions on the Web (TWEB)* 13, 1 (2018), 1–27.

[55] Justin Woodring, Katherine Perez, and Aisha Ali-Gombe. 2024. Enhancing privacy policy comprehension through privacify: A user-centric approach using advanced language models. *Computers & Security* 145 (2024), 103997.

[56] Yutao Wu, Kun Ran, Ming Liu, Adam PA Cardilini, Arif Nurwidyantoro, and Xiao Liu. 2025. CLEAR: Climate Policy Retrieval and Summarization Using LLMs. In *Companion Proceedings of the ACM on Web Conference 2025*. 2927–2930.

[57] Le Yu, Xiapu Luo, Xule Liu, and Tao Zhang. 2016. Can we trust the privacy policies of android apps?. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 538–549.

[58] Razieh Nokhbeh Zaeem, Rachel L German, and K Suzanne Barber. 2018. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology (TOIT)* 18, 4 (2018), 1–18.

[59] Shikun Zhang, Yuanyuan Feng, Yaxing Yao, Lorrie Faith Cranor, and Norman Sadeh. 2022. How Usable Are iOS App Privacy Labels? *Proceedings on Privacy Enhancing Technologies* 2022, 4 (2022), 204–228. https://doi.org/10.56553/popets-2022-0106

[60] Shikun Zhang, Lily Klucinec, Kyerra Norton, Norman Sadeh, and Lorrie Faith Cranor. 2024. Exploring {Expandable-Grid} Designs to Make {iOS} App Privacy Labels More Usable. In *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*. 139–157.

[61] Shikun Zhang and Norman Sadeh. 2023. Do Privacy Labels Answer Users' Privacy Questions?. In *Proceedings of the Symposium on Usable Security and Privacy (USEC)*. Internet Society. https://doi.org/10.14722/usec.2023.232482

[62] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proceedings on privacy enhancing technologies* (2019).

A Data Safety Label Example

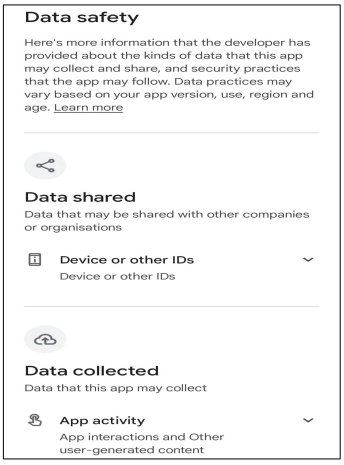

Figure 11: An Example of Data Safety Label

B Data Safety Label Data Structure Example

Listing 1: Data Safety label for the Bluesky app

```
{
  "Bluesky": {
    "Data shared": {
      "App info and performance": "Crash logs, Diagnostics,
     and Other app performance data",
      "Device or other IDs": "Device or other IDs",
      "Personal info": "User IDs",
      "App activity": "App interactions and Other user-
     generated content"
    },
    "Data collected": {
      "App info and performance": "Crash logs, Diagnostics,
     and Other app performance data",
      "Device or other IDs": "Device or other IDs",
      "Photos and videos": "Photos",
      "Personal info": "Email address and User IDs",
      "App activity": "App interactions and Other user-
     generated content"
    }
  }
}
```

C Privacy Policy Data Structure Example

Listing 2: Data types collected and shared with third parties according to the Bluesky privacy policy

```
{
  "Bluesky": {
    "collected": [
      "email address",
      "phone number",
      "images you associate with your profile",
      "birth date",
      "username",
      "name",
      "posts",
      "comments",
      "direct messages",
      "any personal information you provide with your
     application",
      "payment information",
      "card information",
      "Internet protocol (IP) address",
      "user settings",
      "cookie identifiers",
      "mobile carrier",
      "other unique identifiers",
      "browser or device information",
      "Internet service provider (ISP)",
      "the posts you view on Bluesky",
      "the links you click within Bluesky",
      "the frequency and duration of your activities",
      "password",
      "birthday",
      "telephone number",
      "government identifiers",
      "Age",
      "address",
      "passport",
      "account information",
      "driver's licence number",
      "passport number",
      "partial debit card information",
      "partial credit card information",
      "bank account information",
      "other payment or financial information",
      "date of birth",
      "gender",
      "purchase history",
      "device information",
      "log data",
      "pictures and videos (content) you upload",
      "role with the business or organisation",
      "communication preferences",
      "chat history",
      "messages you send us",
      "generated content"
    ],
    "shared": [
      "email address",
      "phone number",
      "profile image",
      "birth date",
      "username",
      "name",
      "posts",
      "comments",
      "direct messages",
      "payment information",
      "card information",
      "Internet protocol (IP) address",
      "user settings",
      "cookie identifiers",
      "mobile carrier",
      "other unique identifiers",
      "browser or device information",
      "Internet service provider (ISP)",
```

```
      "the posts you view on Bluesky",
      "the links you click within Bluesky",
      "the frequency and duration of your activities",
      "password",
      "birthday",
      "telephone number",
      "government identifiers",
      "age",
      "address",
      "passport",
      "account information",
      "driver's licence number",
      "passport number",
      "partial debit card information",
      "partial credit card information",
      "bank account information",
      "other payment or financial information",
      "date of birth",
      "gender",
      "purchase history",
      "device information",
      "log data",
      "pictures and videos (content) you upload",
      "role with the business or organisation",
      "communication preferences",
      "chat history",
      "messages you send us",
      "generated content"
    ]
  }
}
```

D SRS Calculation in Practice

To illustrate how the SRS is calculated in practice, we provide an example for a single app under the proposed weighting scheme. Consider an app x with three data categories $C_x = \{i, j, k\}$. Assume sensitivity weights are $w_i = 3$ (high), $w_j = 2$ (medium), and $w_k = 1$ (low). For the *collection* operation, suppose the policy and label agree on categories j and k but disagree on i. Then $\text{Cons}_{x,i}^{(\text{collect})} = 0$ and $\text{Cons}_{x,j}^{(\text{collect})} = \text{Cons}_{x,k}^{(\text{collect})} = 1$. By Eq. (11),

$$\text{SRS}_x^{(\text{collect})} = \frac{3(1-0) + 2(1-1) + 1(1-1)}{3+2+1} = \frac{3}{6} = 0.50.$$

For *sharing*, assume the app disagrees on i and j but agrees on k, so $\text{Cons}_{x,i}^{(\text{share})} = 0$, $\text{Cons}_{x,j}^{(\text{share})} = 0$, and $\text{Cons}_{x,k}^{(\text{share})} = 1$. Then

$$\text{SRS}_x^{(\text{share})} = \frac{3(1-0) + 2(1-0) + 1(1-1)}{6} = \frac{5}{6} \approx 0.83.$$

The unweighted overall score (Eq. (12)) is $\text{SRS}_x^{(\text{overall})} = (0.50 + 0.83)/2 \approx 0.67$. Using the weighted overall score (Eq. (13)) with $\alpha = 0.6$,

$$\text{SRS}_x^{(\text{overall-w})} = 0.6(0.83) + 0.4(0.50) \approx 0.70,$$

placing app x at the boundary of the *high-risk* tier under our thresholds.

E Additional Quantitative Analysis

RQ5. How does privacy misalignment risk relate to app popularity (user ratings and installs)?

To examine whether privacy-sensitive disclosure risk is associated with app popularity, we correlate the overall weighted Sensitivity Risk Score (SRS) with app rating and download count. Figure 12 plots the overall weighted score $\text{SRS}_i^{(\text{overall-w})}$ against the app's average Google Play rating. The points form a broad cloud rather than a clear trend: across the common rating range of 3.0–4.8, we observe a full spectrum of risk scores, from near 0 to above 0.8. In particular, many highly rated apps (4+ stars) exhibit elevated privacy-sensitive risk scores, indicating substantial misalignment in the disclosure of sensitive data categories. Visually, there is no clear evidence that higher user ratings are associated with lower SRS values. This suggests that privacy-sensitive disclosure risk is either not readily observable to users or does not systematically influence their rating behavior.

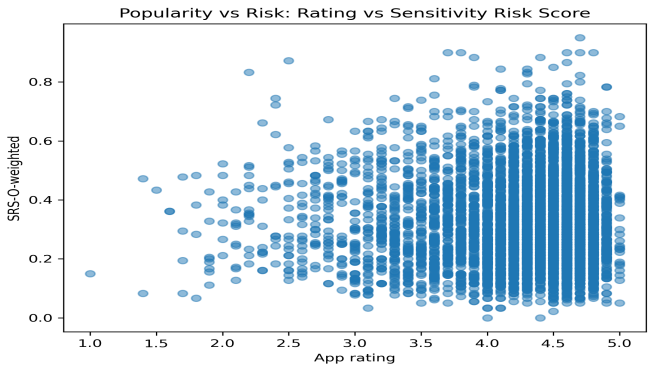

Figure 12: Rating Vs Risk Score

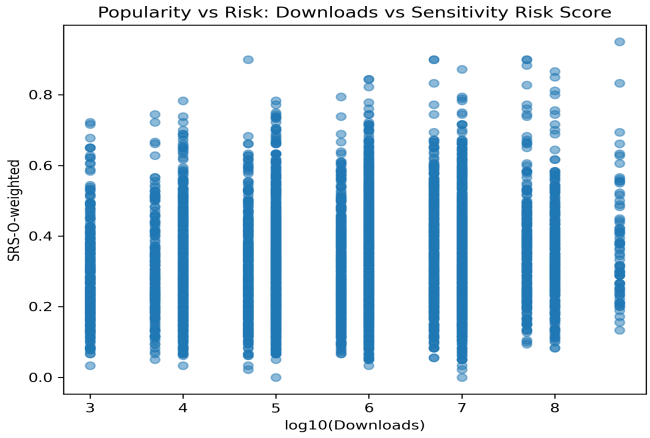

Figure 13: Download Vs Risk Score

Figure 13 shows a similar analysis for app downloads, using $\log_{10}$ of the install count. Each vertical band corresponds to a download bracket (e.g., 10^3, 10^4, …, 10^8 installs). Within each bracket, the distribution of risk scores is broadly similar, and high SRS values appear even among the most downloaded apps. We therefore find no clear pattern indicating that more widely installed apps are

less prone to PPD–DSL misalignment. Overall, misalignment risk appears largely independent of conventional popularity signals, such as ratings and download counts.

RQ5 Key Findings

Overall, the results suggest that the privacy-sensitive risk score is largely independent of app popularity as measured by user ratings and download counts. High-risk apps appear across rating levels and install brackets, indicating that widely used or highly rated apps are not necessarily lower risk from a privacy-disclosure perspective.

F Additional Qualitative Analysis

F.1 Case Study: OpenAI, "No Data Shared" Labels, and Third-Party Analytics

The ChatGPT Android app on Google Play reports a narrow view of sharing in its Data Safety label. In the "Data shared" section, the listing states that the app "may share" only *Device or other IDs* with third parties, for purposes such as advertising or marketing and fraud prevention, security, and compliance, while collected sections indicate that the app collects location, personal information, app info and performance, messages, and app activity data [38]. To an ordinary user, this combination suggests that only a single, relatively abstract identifier leaves their device and that richer information, such as their name, email address, or usage details, is not shared with other companies. OpenAI's own privacy policy explains that OpenAI collects multiple categories of personal information including account information (name, contact details, payment data), user content, device and log data, and information obtained via cookies and similar technologies and may share this data with a range of third-party service providers such as cloud hosting, analytics, security, and customer-support vendors, as well as with affiliates and, in some circumstances, other third parties [39]. While these transfers are framed as processing "on OpenAI's behalf," from a user perspective they still represent disclosure of data to external entities that can see and process their information. The recent Mixpanel incident makes these abstract disclosures concrete. In November 2025, OpenAI reported that a security breach at Mixpanel, a web analytics provider used on the frontend interface for its API product, resulted in the export of a dataset containing *limited analytics data* for some API users [40]. OpenAI specified that the exposed fields included names, email addresses, approximate location based on browser activity, browser and operating system details, referring websites, and organization or user IDs associated with API accounts, while emphasizing that no chat content, API requests, passwords, API keys, payment details, or government IDs were compromised and that ChatGPT end-users were not affected [40]. Even though this breach occurred in Mixpanel's systems rather than OpenAI's, it illustrates the concrete types of personal and device data that third-party analytics partners can hold in practice. Overall, this case study highlight the limitations of high-level app-store labels for communicating third-party data flows. On the user side, a label that only names "Device or other IDs" as shared does not help users understand that identified or identifiable metadata (such as names, email addresses, locations, and detailed telemetry) may be accessible to analytics vendors under the broader privacy policy. On the developer side, OpenAI is operating within the current rules: programmatic analytics and telemetry can be classified as "service provider" processing, so only a minimal shared-data category appears in the label, while the detailed description of what those vendors actually receive surfaces only in the privacy policy or, in the worst case, in a breach notification.

G Ablation Analysis on Generic Sharing Statements

To evaluate the sensitivity of our findings to generic sharing language, we conducted a proportional ablation analysis targeting privacy policies containing vague sharing statements such as "we may share your information," "we may disclose your information," "shared with partners," "shared with affiliates," or "shared with third parties" without explicitly specifying the associated shared data categories. Across the 6,051 analyzed apps, we identified 1,315 apps containing such generic sharing statements, corresponding to 21.73% of the analyzed corpus. Since this ablation only affects sharing-related privacy-policy disclosures, all collection-related metrics and all Data Safety Label values remain unchanged.

Sharing-related metrics were adjusted using:

$$M'_{share} = (1 - r) \cdot M_{share} \tag{14}$$

where M_{share} denotes the original sharing-related metric, M'_{share} denotes the adjusted metric after the ablation, and:

$$r = \frac{1315}{6051} = 21.73\% \tag{15}$$

represents the fraction of apps containing generic sharing statements. For SRS-related metrics, the reported values correspond to the mean app-level SRS scores across the analyzed corpus.

Table 4 summarizes the resulting changes. While the ablation reduces sharing-related inconsistency metrics, substantial disclosure misalignment remains even after masking generic sharing statements. This confirms that the observed inconsistencies are not solely driven by the generic-sharing inference rule and that the main findings observed in the study remain consistent in this sensitivity analysis.

Table 4: Proportional ablation analysis on generic sharing statements.

Metric	Original	After Ablation	Difference
Collection Consistency	66.9%	66.9%	0.0%
Sharing Consistency	68.9%	75.7%	+6.8%
DSL-only Sharing Misalignment	26.0%	20.4%	-5.6%
PPD-only Sharing Misalignment	5.0%	3.9%	-1.1%
Collection Cohen's κ	31.0%	31.0%	0.0%
Sharing Cohen's κ	16.0%	19.5%	+3.5%
$SRS^{collect}$	33.0%	33.0%	0.0%
SRS^{share}	31.3%	24.5%	-6.8%
$SRS^{overall}$	32.2%	28.8%	-3.4%
$SRS^{overall-w}$ ($\alpha = 0.6$)	32.0%	27.9%	-4.1%

H Robustness of SRS-based risk stratification.

Since the weighted overall SRS uses α to control the relative emphasis on sharing versus collection, we further examine whether the risk-tier distribution remains stable when this parameter varies. In addition to the baseline setting of $\alpha = 0.6$, we evaluate alternative settings of $\alpha \in \{0.4, 0.5, 0.7\}$ while keeping the same low, medium, and high tier thresholds.

As shown in Table 5, the distribution remains consistent across these settings. The proportion of high-risk applications stays within a narrow range of 1.21%–1.37%, while the low- and medium-risk tiers show only modest variation. Moreover, 90.56%–96.2% of applications remain in the same risk tier as in the baseline setting, indicating that the observed stratification is stable under reasonable changes to α. This stability suggests that the RQ3 findings are not driven by the specific baseline choice of $\alpha = 0.6$. Rather, the low, medium, and high risk patterns reflect consistent disclosure-misalignment structure across the analyzed applications.

Table 5: Sensitivity analysis of risk-tier distribution under alternative α values.

α	Low (%)	Medium (%)	High (%)	Same Tier (%)
0.4	46.84	51.96	1.21	90.56
0.5	49.96	48.83	1.21	96.00
0.6	49.61	49.10	1.29	100.00
0.7	51.02	47.61	1.37	96.20

I Emergent Data Types Not Captured by Official Data Safety Categories

Our analysis of non-schematized policy terms reveals several recurring data types that do not map cleanly onto any of Google Play's official Data Safety Label data types. In particular, privacy policies frequently mention items such as *login credentials account passwords*, *search history* outside the app context, *operating system, IP address and device type*, and highly sensitive identifiers like *Social Security numbers*, *passport numbers*, or *fingerprints*. These appear as distinct, privacy relevant data fields in policy text, yet they either lack a dedicated category in the Data Safety schema or are only indirectly represented through broader labels as "Others".

As a concrete example, AccuWeather: Weather Radar's privacy policy states that when users create an account or sign in, the service collects "profile, credentials, name, username, email address, and password ('Login Information')" and generates account identifiers and technical logs to administer and protect the account [1]. However, Google Play's Data Safety taxonomy does not expose "password" as a separate data type; instead, these authentication flows are folded into personal-info types such as *Email address* and *User IDs* and purposes like *Account management* and *Security* [20]. This mismatch illustrates a broader gap between the granularity of data types described in privacy policies and the coarser set of categories that are ultimately disclosed to users in the Data Safety section.