# A Large-Scale Privacy Measurement of Android SDKs

*Mark Huasong Meng (I2R/NUS), Chuan Yan (UQ), Qing Zhang (ByteDance), Kailong Wang (HUST), Sin Gee Teo (I2R), Guangdong Bai (UQ), and Jin Song Dong (NUS)*

*SinFra 2024, Singapore, June 24-25*

Third-party Software Development Kits (SDKs) are widely adopted in Android app development, to effortlessly accelerate development pipelines and enhance app functionality. However, this convenience raises substantial concerns about unauthorized access to users' privacy-sensitive information, which could be further abused for illegitimate purposes like user tracking or monetization. Our study offers a targeted analysis of user privacy protection among Android third-party SDKs, filling a critical gap in the Android software supply chain. It focuses on two aspects of their privacy practices, including *data exfiltration* and *behavior-policy compliance* (or *privacy compliance*), utilizing techniques of taint analysis and large language models. It covers 158 widely-used SDKs from two key SDK release platforms, the official one and a large alternative one. From them, we identified 338 instances of privacy data exfiltration. On the privacy compliance, our study reveals that more than 30% of the examined SDKs fail to provide a privacy policy to disclose their data handling practices. We also find that 37% of them over-collect user data, and 88% falsely claim access to sensitive data.
Keywords: mobile security, security measurement, AI/LLM for security, static taint analysis.

## Problem Definitions and Preliminaries

This work aims to (1) detect potential exfiltration of users' privacy data existing in the third-party SDKs, (2) explore the privacy compliance of developers in the SDK release, and thereby, (3) unveil the landscape of privacy protection in the Android ecosystem at the SDK level. To this end, we propose a compliance model to systematically assess the privacy practice of Android SDK regarding personal information collection. We define three types of compliance issues concerned in our investigation.

**Type I: Privacy Leakage**. We define that any pre-recognized privacy data $d$ being observed in sharing operations of an SDK $s$ would constitute privacy sharing risk, detailed as $\exists s \in \mathbb{S}, \exists d \in \mathbb{D}, d \in \mathcal{U}_s \Rightarrow \nvdash_{\mathcal{P}} \mathcal{C}_s$.

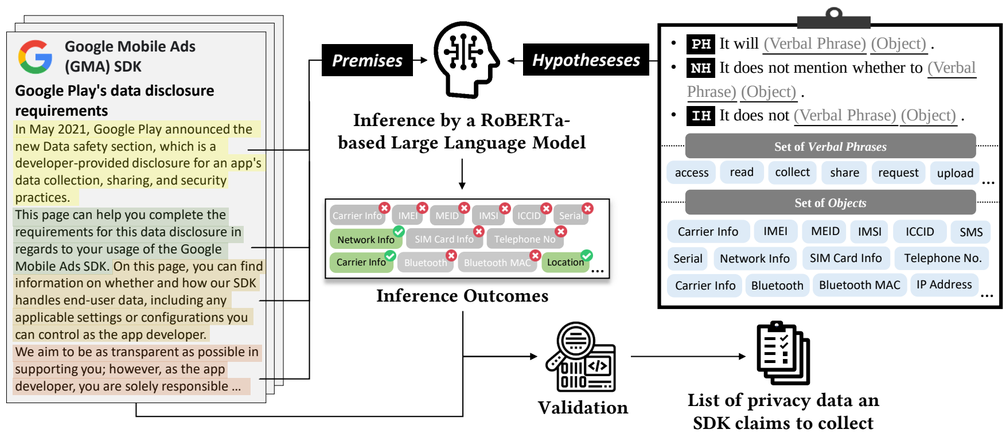
**Type II: Excessive Collection**. Excessive collection concerns privacy practice of an SDK that attempts to read more types of privacy data than what it requests/claims in its privacy policy. Excessive collection infringes on users' *right to be informed* as the SDK collects privacy data without users' awareness and consent. We define this type of risk as: $\exists s \in \mathbb{S}, \exists d \in \mathbb{D}, (d \in \mathcal{R}_s) \wedge (d \notin \mathcal{C}_s) \Rightarrow \mathcal{R}_s \nvdash_{\mathcal{P}} \mathcal{C}_s$.

**Type III: Over-claiming**. We assert an SDK that has an over-claiming issue if it is found to claim its access to more types of privacy data than what it actually reads. Over-claiming issues, although may not substantially lead to privacy data exfiltration, seriously violate the *data minimization principle* stipulated in Article 5(1)(c) of GDPR. It is defined as $\exists s \in \mathbb{S}, \exists d \in \mathbb{D}, (d \notin \mathcal{R}_s) \wedge (d \in \mathcal{C}_s) \Rightarrow \mathcal{R}_s \nvdash_{\mathcal{P}} \mathcal{C}_s$.

**Collecting Android SDKs**. We managed to collect **158** unique SDKs from Google Play SDK Index and CAICT website as of October 2022, which constitute $\mathbb{S}$ in the problem definitions.

**Scope of Privacy Data**. We resort to the Android developer documentation to define the list of privacy data. In addition, we also take existing literature in personal information collection and app analysis into consideration to complement the list of privacy data. As a result, we managed to recognize **41** types of privacy data including IMEI, serial, Google Ad IDs, etc. These 41 types of privacy data constitute $\mathbb{D}$ in the problem definitions.

## Data Collection Inference by LLM




The workflow of our privacy policy analysis to find out what an SDK claims to collect

**Adoption of Natural Language Inference (NLI) Model**. In this work, we leverage the "`roberta-NLI`" model, a state-of-the-art transformer-based masked language model specialized in NLI tasks, to help us identify data requested from the collected privacy policies.

**Inference Setup**. To avoid the disturbance of irrelevant context, we tailor our hypotheses for each data type Instead of a single hypothesis to determine if the premise *mentions* the collection of certain data, we devise three hypotheses corresponding to three different sentiments to regulate the contextual scope, namely *positive hypothesis* (shown as **PH** in the figure above, e.g., "It will collect ..."), *negative hypothesis* (**NH**, e.g., "It does not collect ..."), and *irrelevant hypothesis* (**IR**, "It does not mention whether to collect ...").

This work is currently under peer-review. Subsequent research on AI/LLM for software engineering/security has been published in PETS, ICSE, and FSE.

## Featured Findings

Evaluation outcomes of LLM-based privacy policy analysis

| Data type & class | #SDKs detected to collect | TP | FP | Precision | Data type & class | #SDKs detected to collect | TP | FP | Precision |
|---|---|---|---|---|---|---|---|---|---|
| **C1** Carrier info | 12 | 10 | 2 | 83.3% | **C3** Location | 28 | 25 | 3 | 89.3% |
| Device identifiers* | 21 | 21 | 0 | 100.0% | Misc sensors | 1 | 1 | 0 | 100.0% |
| Network info | 30 | 25 | 5 | 83.3% | Android ID | 8 | 8 | 0 | 100.0% |
| SIM card info | 7 | 7 | 0 | 100.0% | App list | 4 | 4 | 0 | 100.0% |
| SMS | 7 | 3 | 4 | 42.9% | Audio record | 10 | 10 | 0 | 100.0% |
| Telephone No | 5 | 3 | 2 | 60.0% | Google Ad ID | 6 | 6 | 0 | 100.0% |
| **C2** BSSID/SSID | 4 | 4 | 0 | 100.0% | OAID | 2 | 2 | 0 | 100.0% |
| IP address | 47 | 45 | 2 | 95.7% | **C4** Account info | 10 | 4 | 6 | 40.0% |
| WiFi | 4 | 3 | 1 | 75.0% | Clipboard | 1 | 1 | 0 | 100.0% |
| **C3** Camera | 7 | 5 | 2 | 71.4% | **Total** | **214** | **187** | **27** | **87.4%** |

Assessment of privacy compliance by data types

| Data type and class | Type II (EC) | Type III (OC) | Data type and class | Type II (EC) | Type III (OC) | Data type and class | Type II (EC) | Type III (OC) |
|---|---|---|---|---|---|---|---|---|
| **C1** Carrier info | 10 | 7 | **C2** Bluetooth | 4 | 0 | **C4** Android ID | 10 | 0 |
| IMEI | 1 | 6 | Bluetooth MAC | 2 | 0 | App list | 22 | 1 |
| MEID | 0 | 21 | BSSID/SSID | 8 | 2 | Audio record | 0 | 9 |
| IMSI | 1 | 18 | IP Address | 2 | 44 | Google Ad ID | 2 | 3 |
| ICCID | 4 | 16 | WiFi | 3 | 0 | OAID | 0 | 3 |
| Network info | 0 | 24 | WiFi MAC | 2 | 2 | Screen record | 0 | 1 |
| Serial | 1 | 6 | *Subtotal* | *(21)* | *(48)* | *Subtotal* | *(34)* | *(17)* |
| SIM card info | 1 | 0 | **C3** Camera | 0 | 4 | **C5** Account info | 1 | 4 |
| SMS | 3 | 2 | Location | 11 | 14 | Clipboard data | 11 | 0 |
| Telephone no | 1 | 3 | Misc. sensors | 7 | 0 | Contact list | 1 | 10 |
| *Subtotal* | *(22)* | *(103)* | *Subtotal* | *(18)* | *(18)* | *Subtotal* | *(13)* | *(14)* |
| | | | | | | **Total** | **108** | **200** |

## Conclusion

Our findings were startling, with 338 potential privacy leakages detected from 158 SDKs. Our study reveals that less than 70% of examined SDKs provide privacy policies, among which approximately 37% of SDKs are found indulging in the over-collection of privacy data, signaling a clear violation of privacy norms. We believe that tighter regulatory measures, improved development guidelines, and transparent permission management are essential steps toward mitigating privacy risks in Android applications.