Multilingual Compliance: A Comparative Study of Privacy Policies in Chinese, Japanese, and Korean

Muhammad Hassan
University of Illinois Urbana-Champaign
Champaign, IL, 61820
mhassa42@illinois.edu

Phillip K Nakamurar
University of Illinois Urbana-Champaign
Champaign, IL, 61820
pnaka2@illinois.edu

Yuanye Ma
Discovery Partners Institute, University of Illinois
Chicago, IL, 60606
yuanyem@uillinois.edu

Abstract

As global data privacy regulations tighten, privacy policies have become pivotal in shaping the relationship between users and online services. Despite their importance, these documents remain inaccessible to many users due to excessive length and complex legal terminology, posing significant challenges to informed consent. Moreover, most existing research has concentrated on English-language policies, leaving a gap in our understanding of how privacy policies are structured and enforced in non-English-speaking regions.

To address this, we present a study that examines privacy policies from Chinese, Japanese, and Korean websites, analyzing their compliance with respective national privacy laws. We collected and processed a dataset of over 2,400 privacy policies, and more than 5.2 million tokens, employing language detection and text analysis techniques to assess adherence to regulatory frameworks, including China's PIPL, Japan's APPI, and Korea's PIPA. Our findings reveal notable disparities in regulatory compliance, with Chinese and Japanese policies generally demonstrating stronger alignment with legal standards compared to Korean policies.

These results highlight the need for enhanced clarity and enforcement across regions, offering insights into the evolving global landscape of data privacy. Our work underscores the importance of multilingual analysis in advancing the accessibility and transparency of privacy practices, laying the groundwork for future research and policy development in this critical domain.

1 Introduction

The widespread deployment of machine learning (ML) technologies has heightened concerns about how user data is collected, processed, and governed. Privacy policies serve as the legal framework by which websites and digital platforms disclose their data practices to users, forming a critical part of data governance strategies. These policies, rooted in the *notice and choice* model, provide the basis for websites to outline their data handling procedures, enabling users to make informed decisions about their privacy. With the increasing focus on algorithmic transparency and accountability, privacy policies have emerged as essential tools for ensuring compliance with evolving data protection regulations as identified byChang et al. [2018]. Zaeem and Barber [2020], Bonta [2022] claim that regulatory frameworks such as the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and Brazil's Lei Geral de Proteção de Dados (LGPD) have raised the bar for

2nd Workshop on Regulatable ML at NeurIPS 2024.

transparency, mandating stringent requirements for data collection, sharing, and user consent Gupta et al. [2024].

Despite the global expansion of privacy regulations, a significant gap remains in ensuring that privacy policies and machine learning (ML) systems align with evolving legal standards for data practices. Existing privacy laws generally do not specify precise operational requirements for ML systems, leaving privacy policies as the primary mechanism to communicate data usage practices. As such, privacy policies serve as a de facto regulatory proxy, bridging the gap between general privacy laws and specific operational practices in ML and data-driven applications. This relationship — where legal frameworks set broad protections, ML systems need specific compliance guidance, and privacy policies serve as the main point of regulatory communication — presents a complex challenge for privacy protection focused on user needs. Understanding this dynamic is crucial, as it frames privacy policies not just as legal documents but as key safeguards that connect regulatory standards with the practices of ML systems.

While there has been considerable research on privacy policy text analysis, much of this work has been confined to English-language regulations, leading to a limited understanding of how different regulatory environments impact ML model compliance across non-English-speaking regions. To bridge this gap, our work investigates privacy policies from three major non-English languages—Chinese, Japanese, and Korean—each of which is governed by distinct regulatory frameworks, including China's Personal Information Protection Law (PIPL), Japan's Act on the Protection of Personal Information (APPI), and Korea's Personal Information Protection Act (PIPA). By analyzing these policies using natural language processing (NLP) techniques, we aim to provide empirical evidence of the operational challenges that arise when translating diverse regulatory principles into algorithmic practices. Our study assesses compliance while also highlighting tensions between regulatory desiderata like privacy, fairness, and transparencyas they relate to machine learning, offering direction for future research.

To summarize, this paper presents a multilingual privacy policy corpus and a preliminary cross-lingual, cross-jurisdictional analysis, highlighting operational gaps between privacy regulations and ML practices; our findings underscore the need for regulatory-compliant frameworks, particularly in non-English regions, and establish a foundation for future research addressing tensions between compliance and ML system design.

2 Related Work

Several studies have explored the challenges of ensuring compliance with privacy laws, particularly the GDPR and CCPA, within the context of algorithmic decision-making systems. These studies highlight that while regulations emphasize transparency, fairness, and accountability, implementing these principles in practice often presents significant obstacles for organizations deploying ML models Goldsteen et al. [2022], Murakonda and Shokri [2020]. Recent large-scale corpus analyses of privacy policies, including works like Amos et al. [2021], Belcheva et al. [2023], Srinath et al. [2021], Wagner [2023], Wilson et al. [2016], Nokhbeh Zaeem and Barber [2021], have contributed substantial datasets and longitudinal analyses, advancing the understanding of policy compliance trends across time and regions.

Text analysis and natural language processing (NLP) techniques have been increasingly used to analyze privacy policies, offering insights into compliance levels across various sectors and jurisdictions. Prior work has focused primarily on English-language privacy policies, using computational methods to detect non-compliance, assess legal ambiguities, and identify thematic trends within policy documents Papanikolaou et al. [2011], Ravichander et al. [2021], Del Alamo et al. [2022]. These studies provide valuable methodologies for policy analysis, yet they fall short in addressing multilingual and multi-jurisdictional complexities, which are critical in global ML deployments. Research into non-English privacy policies remains limited, leading to a skewed understanding of regulatory compliance on a global scale. Some notable exceptions include Arora et al. [2022], Mashaabi et al. [2023], which have analyzed bilingual corpora in German and Arabic privacy policies, respectively, highlighting regional linguistic and regulatory diversity in privacy policies.

Research increasingly explores how different regulatory frameworks intersect with machine learning practices, particularly in areas like data governance, cross-border data transfers, and consent management Carter [2020]. However, many studies examine these frameworks in isolation, focusing

on individual countries rather than comparing compliance across jurisdictions. This approach often misses the nuanced tensions between key principles, such as data minimization, transparency, and fairness, which can vary significantly between legal systems Coche et al. [2024]. These gaps are especially relevant in non-English-speaking regions, where regulatory standards are distinct yet equally rigorous.

Our work builds on existing methodologies by addressing current linguistic and jurisdictional blind spots in privacy policy analysis. Specifically, we contribute a novel multilingual dataset of privacy policies in Chinese, Japanese, and Korean, enabling detailed cross-lingual and cross-jurisdictional compliance analysis. By examining policy alignment with regional privacy laws, this study provides new insights into the operational challenges organizations face in adhering to these diverse regulatory requirements. This work establishes an empirical foundation for developing region-specific policy analysis frameworks essential to navigating increasingly complex privacy landscapes globally.

3 Methodology

To create our multilingual privacy policy corpus, we utilized the work of Tranco top website collection, a work done by Le Pochat et al. [2019]. We developed a custom configuration to query the top 100,000 domains for 12 period between June 2023 and June 2024 Le Pochat et al. [2019], and collected a $\text{dataset}_{\text{initial}}$ of top 100K domains [1]. Using the domains in the $\text{dataset}_{\text{initial}}$, we visited each domain for language detection of website. After removing the HTML tags, we used `langdetect` framework to label language of website for each domain in $\text{dataset}_{\text{initial}}$, which supports 55 ISO-639-1 languages as described by Danilk [2016]. Requests were managed using randomized user agents and proxies to ensure the accuracy of language detection and avoid automated blocking mechanisms. After detecting the language and labeling, domains which had websites in Korean, Chinese and Japanese languages were placed in a different $\text{dataset}_{\text{final}}$.

A recursive crawling process was employed to extract internal links from these identified domains with relevant languages in $\text{dataset}_{\text{final}}$, using a tiered approach limited to three levels to reflect typical user behavior. To optimize efficiency, the crawling was parallelized using `ThreadPoolExecutor`[2], allowing up to 20 threads to process domains simultaneously, with a five-minute cap on each domain's total processing time. This approach helped to balance comprehensive data collection and computational resource management. This helped in collecting privacy policy links for each domain in $\text{dataset}_{\text{final}}$.

Subsequently, Privacy policies were retrieved by filtering internal links for relevant keywords, such as "privacy," "GDPR," and "cookies." The content was then parsed with BeautifulSoup, and hash-based deduplication was applied to prevent redundancy. Boilerplate text was preserved to avoid accidental removal of critical policy content during preprocessing. Each policy's language was verified, and the corresponding metadata was updated in the database for structured storage.

For compliance assessment, a reviewer proficient in privacy research examined translated versions of China's PIPL, Japan's APPI, and South Korea's PIPA to identify key principles, such as "Purpose Limitation," "Data Minimization," and "Content Transparency." These principles served as benchmarks for evaluating compliance. Language-specific tokenization tools—SudachiPy for Japanese, Jieba for Chinese, and Okt for Korean—were used to preprocess the privacy policies. Tokenization and normalization included the removal of stop words, punctuation, and extraneous characters. The preprocessed data was then stored in structured formats for linguistic and compliance analysis.

The text analysis pipeline matched key terms and their synonyms in each privacy policy with the identified regulatory principles. Term presence and frequency were used to generate a compliance score for each policy, with scores normalized between 0 and 1. Detailed score visualizations are presented in Appendix A.

4 Results and Discussion

In this section, we present an analysis of privacy policy compliance within the regulatory frameworks of China, Korea, and Japan, using a dataset of 2,438 privacy policies across these three languages.

[1] Available at https://tranco-list.eu/list/J9LYY

[2] https://docs.python.org/3/library/concurrent.futures.html

The dataset includes 1,296 Japanese policies, 698 Korean policies, and 444 Chinese policies, resulting in over 5 million tokens. Although Japanese websites dominate the dataset due to their prominence in the Tranco ranking, the overall token distribution across languages is relatively balanced. This robust dataset allows for generalizable insights into the regulatory landscape of each jurisdiction.

Our analysis evaluates compliance with China's Personal Information Protection Law (PIPL), Korea's Personal Information Protection Act (PIPA), and Japan's Act on the Protection of Personal Information (APPI). Privacy policies were analyzed using string-matching techniques following preprocessing, based on key principles identified within each regulatory framework, e.g. user rights, data processing, third-party sharing, and data minimization.

Compliance with China's PIPL Law: Policies from Chinese websites showed strong compliance with core PIPL principles, particularly in *User Rights*, *Purpose Limitation*, and *Data Processing*. However, adherence to *Cross-Border Data Transfers* was lower than expected, despite China's emphasis on data sovereignty. This discrepancy highlights a global trend towards stricter controls on international data flows, which requires further investigation Li [2024].

Compliance with Korea's PIPA Law: The analysis of Korean policies revealed mixed results. While there was reasonable compliance with *Data Security* and *Data Transfer* requirements, areas such as *Third-Party Data Sharing* and *Data Breach Notification* showed lower compliance. This may reflect the ongoing integration of recent PIPA amendments into corporate policies, particularly concerning the appointment of a *Data Protection Officer* and compliance with international data transfer obligations.

Compliance with Japan's APPI Law: Japanese policies displayed high compliance with APPI requirements, particularly in *Purpose of Use*, *Data Minimization*, and *Consent*. However, the legal basis for *Data Processing* was not consistently addressed, indicating potential gaps in how organizations justify their data practices.

Key Findings: Overall, Chinese and Japanese privacy policies exhibited stronger compliance with their respective regulations compared to Korean policies. Stricter enforcement mechanisms in China and Japan, along with more established regulatory frameworks, likely contribute to this trend. It is important to note that our initial review was conducted by a single reviewer for each language, which may limit the granularity of the compliance assessment. Future research can extend this analysis by incorporating more reviewers and exploring the enforcement of these regulations in practice.

Our findings indicate that while regulatory frameworks in these regions are advancing, gaps in compliance persist, particularly in areas of cross-border data transfers and justification of data processing practices. These results align with prior research on the challenges of aligning privacy policy content with evolving legal standards Mulder and Tudorica [2019], Casino et al. [2022]. The visualization of compliance distributions is provided in Figures 1, 2, and 3 in the appendix.

5 Limitations & Future Work

While this study provides a preliminary analysis of privacy policy compliance across multiple languages, there are several areas for further enhancement. One of our primary objectives moving forward is to improve the accuracy of language-based policy identification by expanding collaborations with linguists and regional experts. This will allow us to refine our language detection models and provide more precise insights into policies from diverse linguistic regions.

Additionally, we aim to build a more systematic and scalable pipeline for policy analysis. This will enable a more granular examination of privacy policies, allowing for direct comparisons across regulatory frameworks and regions. By integrating advanced natural language processing techniques, we hope to further assess compliance with emerging privacy regulations and evaluate real-world enforcement.

We anticipate that the insights gained from this workshop will provide valuable feedback and foster new collaborations. These contributions will enhance the depth of our ongoing research and inform future studies on the intersection of privacy, AI, and regulation.

6 Conclusion

In this paper, we examined privacy policy compliance across Chinese, Japanese, and Korean regulatory frameworks. Using a dataset of over 2,400 privacy policies, we applied a methodology that integrates web scraping, language detection, and text analysis to assess alignment with national privacy laws. Findings indicate varying compliance levels, with Japanese and Chinese policies generally more aligned with their respective regulations than Korean policies. These results emphasize the need for enhanced regulatory adherence and identify opportunities for further research into cross-jurisdictional policy analysis. Future work will refine our approach and incorporate insights from academic and regulatory perspectives.

References

Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021*, pages 2165–2176, 2021.

Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar K Bannihatti, Tristan Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, et al. A tale of two regulatory regimes: Creation and analysis of a bilingual privacy policy corpus. In *LREC proceedings*, 2022.

Veronika Belcheva, Tatiana Ermakova, and Benjamin Fabian. Understanding website privacy policies—a longitudinal analysis using natural language processing. *Information*, 14(11):622, 2023.

Rob Bonta. California consumer privacy act (ccpa). *Retrieved from State of California Department of Justice: https://oag. ca. gov/privacy/ccpa*, 2022.

Denise Carter. Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? who is responsible? can the information professional play a role? *Business Information Review*, 37(2):60–68, 2020.

Fran Casino, Claudia Pina, Pablo López-Aguilar, Edgar Batista, Agusti Solanas, and Constantinos Patsakis. Sok: Cross-border criminal investigations and digital evidence. *Journal of Cybersecurity*, 8(1):tyac014, 2022.

Younghoon Chang, Siew Fan Wong, Christian Fernando Libaque-Saenz, and Hwansoo Lee. The role of privacy policy on consumers' perceived privacy. *Government Information Quarterly*, 35(3): 445–459, 2018.

Eugénie Coche, Ans Kolk, and Václav Ocelík. Unravelling cross-country regulatory intricacies of data governance: the relevance of legal insights for digitalization and international business. *Journal of International Business Policy*, 7(1):112–127, 2024.

Michal Danilk. langdetect, Oct 2016. URL `https://pypi.org/project/langdetect/`.

Jose M Del Alamo, Danny S Guaman, Boni García, and Ana Diez. A systematic mapping study on automated analysis of privacy policies. *Computing*, 104(9):2053–2076, 2022.

Abigail Goldsteen, Gilad Ezov, Ron Shmelkin, Micha Moffie, and Ariel Farkash. Data minimization for gdpr compliance in machine learning models. *AI and Ethics*, 2(3):477–491, 2022.

Sonu Gupta, Geetika Gopi, Harish Balaji, Ellen Poplavska, Nora O'Toole, Siddhant Arora, Thomas Norton, Norman Sadeh, and Shomir Wilson. Creation and analysis of an international corpus of privacy laws. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4092–4105, 2024.

Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, NDSS 2019, February 2019. doi: 10.14722/ndss.2019.23386.

Barbara Li. International association of privacy professionals, Apr 2024. URL https://iapp.org/news/a/chinas-new-cross-border-data-transfer-regulations-what-you-need-to-know-and-do.

Malak Mashaabi, Ghadi Al-Yahya, Raghad Alnashwan, and Hend Al-Khalifa. Arabic privacy policy corpus and classification. In *International Conference on Applications of Natural Language to Information Systems*, pages 94–108. Springer, 2023.

Trix Mulder and Melania Tudorica. Privacy policies, cross-border health data and the gdpr. *Information & Communications Technology Law*, 28(3):261–274, 2019.

Sasi Kumar Murakonda and Reza Shokri. Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. *arXiv preprint arXiv:2007.09339*, 2020.

Razieh Nokhbeh Zaeem and K Suzanne Barber. A large publicly available corpus of website privacy policies based on dmoz. In *Proceedings of the eleventh ACM conference on data and application security and privacy*, pages 143–148, 2021.

Nick Papanikolaou, Siani Pearson, and Marco Casassa Mont. Towards natural-language understanding and automated enforcement of privacy rules and regulations in the cloud: survey and bibliography. In *Secure and Trust Computing, Data Management, and Applications: STA 2011 Workshops: IWCS 2011 and STAVE 2011, Loutraki, Greece, June 28-30, 2011. Proceedings 8*, pages 166–173. Springer, 2011.

Abhilasha Ravichander, Alan Black, Tom Norton, Shomir Wilson, and Norman Sadeh. Breaking down walls of text: How can nlp benefit consumer privacy? *Computational linguistics*, 2021.

Mukund Srinath, Shomir Wilson, and C Lee Giles. Privacy at scale: Introducing the privaseer corpus of web privacy policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1, 2021.

Isabel Wagner. Privacy policies across the ages: content of privacy policies 1996–2021. *ACM Transactions on Privacy and Security*, 26(3):1–32, 2023.

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. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, 2016.

Razieh Nokhbeh Zaeem and K Suzanne Barber. The effect of the gdpr on privacy policies: Recent progress and future promise. *ACM Transactions on Management Information Systems (TMIS)*, 12 (1):1–20, 2020.

A Appendix

Figure 1: Chinese Policies

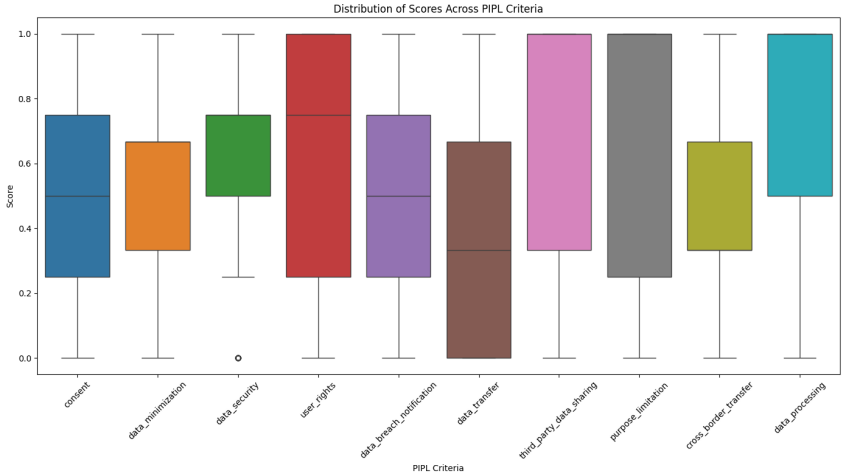


Figure 2: Japanese Policies

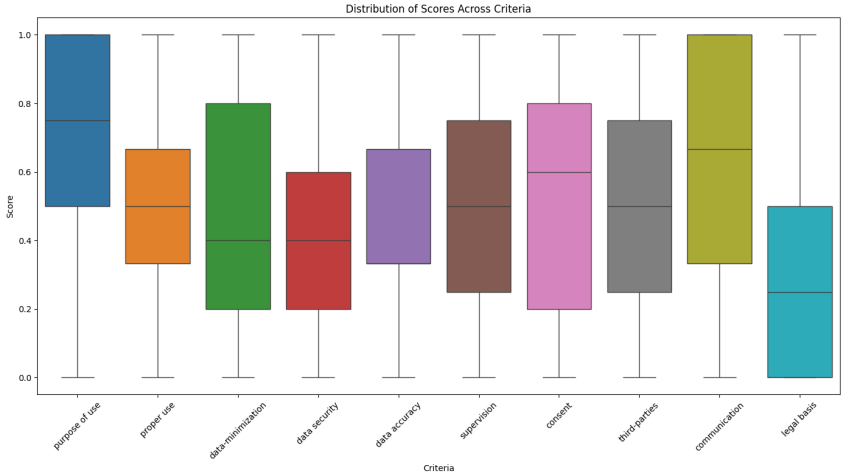
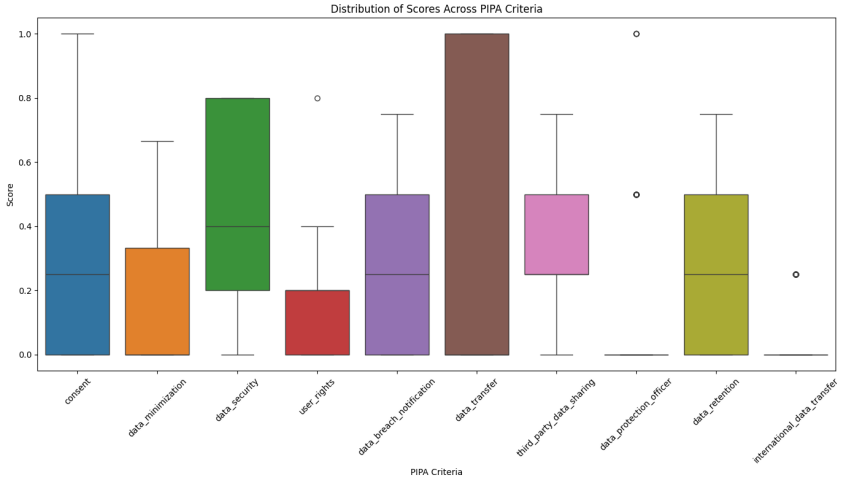


Figure 3: Korean Policies



Category	Count	Tokens
Total Policies	2438	5,245,938
Japanese Policies	1296	1,901,957
Chinese Policies	444	1,433,123
Korean Policies	698	1,910,858

Table 1: Count and Tokens Summary

Language	Average Tokens
Japanese	780.13
Chinese	3,227.75
Korean	2,737.62

Table 2: Average Tokens Summary

Table 3: Full Forms of Privacy Laws

Abbreviation	Full Form
PIPL	Personal Information Protection Law (China)
APPI	Act on the Protection of Personal Information (Japan)
PIPA	Personal Information Protection Act (South Korea)