

APPSI-139: A Parallel Corpus of English Application Privacy Policy Summarization and Interpretation

Pengyun Zhu^{1,2}, Qiheng Sun², Long Wen², Yanbo Wang³, Yang Cao⁴,
Junxu Liu⁵, Deyi Xiong¹, Jinfei Liu^{2,6*}, Zhibo Wang², Kui Ren²

¹ Tianjin University, ² Zhejiang University, ³ North University of China,

⁴ Institute of Science Tokyo, ⁵ The Hong Kong Polytechnic University,

⁶ Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
{pengyunzhu, dyxiong}@tju.edu.cn, jinfeiliu@zju.edu.cn✉

Abstract

Privacy policies are essential for users to understand how service providers handle their personal data. However, these documents are often long and complex, as well as filled with technobabble and legalese, causing users to unknowingly accept terms that may even contradict the law. While summarizing and interpreting these privacy policies is crucial, there is a lack of high-quality English parallel corpus optimized for legal clarity and readability. To address this issue, we introduce APPSI-139, a high-quality English privacy policy corpus meticulously annotated by domain experts, specifically designed for summarization and interpretation tasks. The corpus includes 139 English privacy policies, 15,692 rewritten parallel corpora, and 36,351 fine-grained annotation labels across 11 data practice categories. Concurrently, we propose TCSI-pp-V2, a hybrid privacy policy summarization and interpretation framework that employs an alternating training strategy and coordinates multiple expert modules to effectively balance computational efficiency and accuracy. Experimental results show that the hybrid summarization system built on APPSI-139 corpus and the TCSI-pp-V2 framework outperform large language models, such as GPT-4o and LLaMA-3-70B, in terms of readability and reliability. The source code and dataset are available at <https://github.com/EnlightenedAI/APPSI-139>.

1 Introduction

The privacy policy outlines how service providers collect, process, store, manage, and use the personal information of individuals interacting with their applications (Zhu et al., 2023). Service providers are permitted to handle personal information according to the stipulations outlined in the agreement, with the authorization of users.

However, most privacy policies are often considered “incomprehensible” due to complex technical jargon, legal language, and convoluted grammatical structures (Ermakova et al., 2014; Singh et al., 2011). This issue is further exacerbated by “rational ignorance,” where users perceive the effort required to understand lengthy privacy policies as disproportionate to the benefits, and “dark patterns,” where design elements subtly guide users to quickly click “Agree” or “Join Now” without fully understanding the consequences (Zhu et al., 2018). Such behavior can lead to the disclosure or misuse of users’ sensitive data without explicit consent (Obar and Oeldorf-Hirsch, 2020). Although efforts such as “Privacy Nutrition Labels” (Li et al., 2022; Kollnig et al., 2022), “LPL” (Gerl et al., 2018), and “TILT” (Grünwald and Pallas, 2021) aim to establish standardized and formalized guidelines for privacy policies, assisting service providers in presenting their privacy policies more clearly and understandably, ensuring that service providers genuinely adhere to these guidelines remains a significant challenge.

Automatic summarization is considered a potential solution, utilizing natural language processing and machine learning techniques to automatically extract and summarize key information from text (Keymanesh, 2022; Zayed et al., 2025; Fu et al., 2023). However, due to the lack of parallel corpora for interpreting privacy policies, most existing methods for summarizing privacy policies focus primarily on extracting key information and shortening the length of the policy, while giving less attention to addressing the “incomprehensibility” caused by legal and technical terminology.

In response to the above issues, we have organized domain experts to carefully annotate a parallel corpus for English privacy policy summarization and interpretation, APPSI-139, and based on this, we propose the privacy policy summarization and interpretation framework, Tsci-pp-V2.

*Corresponding Author.

The APPSI-139 corpus includes English privacy policies from 139 top applications across domains such as shopping, live streaming, sports, and gaming. It has been annotated by legal experts to ensure the content is comprehensive, accurate, and easy to understand. The corpus consists of two main components: high-quality annotations and a policy interpretation parallel corpus. Specifically, there are 36,351 annotations covering 11 categories of data practices and 3 special tags, used to identify core clauses, label sensitive information, and recognize potential risks. Additionally, the policy interpretation parallel corpus contains 15,692 pairs, where legal experts have rewritten and paraphrased key clauses containing technobabble and legalese into comprehensible language to reduce the difficulty of understanding.

The TCSI-pp-V2 framework is a Topic-Controlled framework for Summarizing and Interpreting Privacy Policy, built on end-to-end multi-task learning. The framework decomposes the summarization task into five subtasks and employs an alternating training strategy across five parameter-sharing experts, each specialized in one subtask. Specifically, four experts are dedicated to identifying and locating key clauses by detecting their Importance, Risk, Sensitivity, and Topic, while a Rewrite expert is responsible for rewriting and interpreting technobabble and legalese. Through this collaborative architecture, the framework enables precise clause identification and transformation into comprehensible language, supporting users in making informed and cognitively sound privacy decisions.

For clarity, we summarize our main contributions as follows.

- We created APPSI-139, the first parallel corpus of English privacy policies, meticulously annotated and explained by legal experts, aimed at providing user-friendly interpretations to help users make informed privacy decisions.
- We propose the TCSI-pp-V2 framework, which coordinates multiple expert modules through an alternating training strategy, effectively balancing computational efficiency with accuracy.
- Based on APPSI-139 corpus and the TCSI-pp-V2 framework, we develop a summarization and interpretation system. Questionnaire survey evaluations show that, compared to general-purpose models like GPT-4o, our system outper-

forms in terms of readability and reliability, helping users make more informed and rational privacy decisions.

2 Related Work

2.1 Privacy Policy Corpus

In the current research landscape, we observe that publicly available corpus for privacy policies are not only limited in number but also predominantly focused on information extraction or content analysis. For instance, OPP-115 (Wilson et al., 2016) and APP-350 (Zimmeck et al., 2019) are among the earliest known privacy policy corpora, which were compiled using expert annotation and crowd-sourced annotation respectively. These corpora break down privacy policies into finer data practices, providing useful resources for tasks such as classification and content extraction of privacy policy texts. PI-Extract (Bui et al., 2021) is a special project that selects 30 privacy policies from OPP-115 and annotates a more detailed text range for entity recognition tasks. Furthermore, PrivacyQA (Ravichander et al., 2019) and PolicyQA (Ahmad et al., 2020) are privacy policy question-and-answer corpus based on OPP-130. These two corpus consist of pairs made up of the “original text” of the privacy policy and “questions”, which can extract relevant sentences or paragraphs based on the user’s questions. Optoutchoice-2020 (Bannihatti Kumar et al., 2020) and Optoutchoice-2017 (Sathyendra et al., 2017) are two corpora specifically annotated with content related to privacy settings, such as ad tracking, which play an important role in helping users extract and manage privacy settings in privacy policies. Huang et al. (2024) use large language models (LLMs) to create a large-scale classification corpus. The Chinese corpus CA4P-483 (Zhao et al., 2022) was annotated by trained students and includes sentence-level category labels and detailed data processing practices. These privacy policy corpora are mainly designed for extracting specific information or content. CAPP-130 (Zhu et al., 2023) introduced a summary corpus aimed at addressing the “incomprehensibility issue through rewriting, with detailed sentence-level annotations and rewrite examples. However, the English version of CAPP-130 was obtained through machine translation, without adaptation to native English expressions, and lacks the linguis-

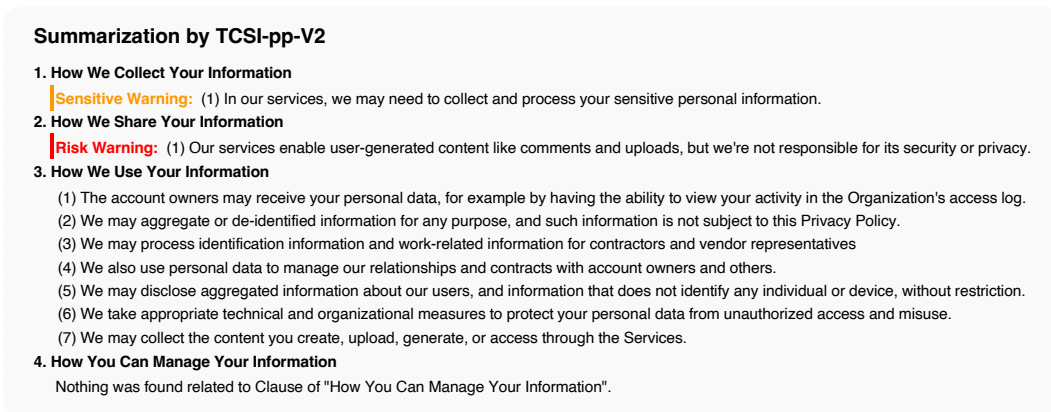


Figure 1: Summarization by TCSI-pp-V2.

tic precision and legal clarity required for effective understanding of privacy policies.

2.2 Privacy Policy Summarization

Recently, natural language processing and machine learning technologies have made significant strides in addressing the readability issues of “length” and “incomprehensibility” in privacy policies. These technologies offer an efficient solution by automatically extracting key information or generating condensed summaries. Specifically, there are three main approaches: extractive summarization, abstractive summarization, and hybrid summarization. *Extractive summarization* improves the readability of privacy policies by selecting sentences or paragraphs related to specific topics. For example, Wilson et al. (2016); Harkous et al. (2018); Liu et al. (2018); Tomuro et al. (2016); Zhao et al. (2022); Ravichander et al. (2019); Ahmad et al. (2020) and PolicyGPT (Tang et al., 2023) tackle policy verbosity by extracting topic-relevant sentences. Meanwhile, Keymanesh et al. (2020); Nokhbeh Zaeem et al. (2022, 2020) focuses on using machine learning to assess and retrieve content that may pose potential privacy risks. *Abstractive summarization* leverages general text simplification techniques (Al-Thanyyan and Azmi, 2021; Alva-Manchego et al., 2020; Siddharthan, 2014) to generate concise and paraphrased summaries. Recent advances in large language models (LLMs)(Saini et al., 2025; Zhu et al., 2024; Li et al., 2024), such as ChatGPT(Ouyang et al., 2022), open new opportunities for privacy policy analysis by simplifying complex linguistic structures (Feng et al., 2023). However, fine-tuned smaller models for specific tasks provide more practical advantages(Qin et al.,

2023), while LLMs remain constrained by computational resources, deployment costs, and the high demands of prompt design when handling lengthy texts (Zhang et al., 2025; Fang et al., 2026; Shi et al., 2024). *Hybrid summarization* combine the strengths of extractive and abstractive approaches, improving the controllability and consistency of generative models (Wang et al., 2017). For example, the TCSI-pp framework (Zhu et al., 2023) outperforms GPT-4 in addressing privacy policy “length” and “incomprehensibility” but still faces computational redundancy due to independent encoding.

3 The APPSI-139 Corpus and TCSI-pp-V2 Framework

In this section, we disclose the details of APPSI-139 in Section 3.1 and present the summarization framework TCSI-pp-V2 in Section 3.2. The summarization examples are shown in Figure 1.

3.1 The APPSI-139 Corpus

Privacy policies often contain a large amount of technical jargon, which requires a strong professional background to fully comprehend. To ensure that the annotations accurately capture the core elements of privacy policies, a team of legal experts including a law professor and five lawyers with masters degrees has developed annotation guidelines for English privacy policy summaries through multiple rounds of pilot annotations. Building on the design of CAPP-130 (Zhu et al., 2023), the guidelines are divided into three parts: *Data Practice Categories*, *Special Markings*, and *Rewritten Sentences*. Figure 2 illustrates the organization of APPSI-139.

Data Practice Categories classify and orga-

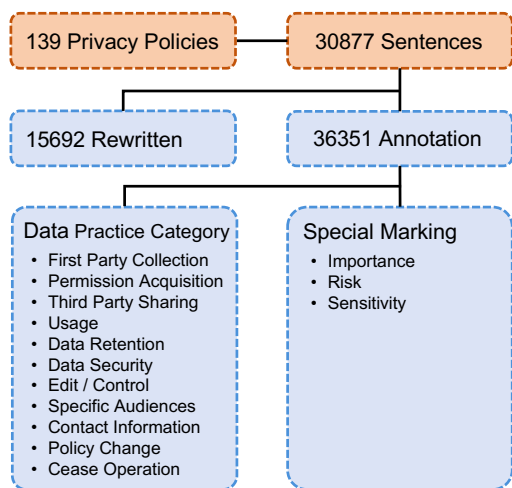


Figure 2: The organization of APPSI-139.

nize information within privacy policies, ensuring clearer and more standardized comprehension of data handling practices. These categories include First Party Collection, Permission Acquisition, Third Party Sharing, Usage, Data Retention, Data Security, Edit/Control, Specific Audiences, Contact Information, Policy Change, and Cease Operation. Detailed definitions of each category are provided in Appendix B.

Special Marking highlights clauses related to important data practices, sensitive personal information, and potential risks, including *Importance*, *Risk*, and *Sensitivity* tags. The definitions are as follows:

- **Importance:** Applies to provisions directly related to key data processing elements, including the definitions of “Risk,” “Sensitivity,” and the eleven categories of “Data Practices” in the clauses. These provisions remain crucial for subsequent summarization and rewriting tasks.
- **Risk:** Applies to content that may raise compliance issues or violate major data protection regulations (e.g., GDPR, CCPA), including vague language about data handling, such as “We may share your location data with third parties without further notice”.
- **Sensitivity:** Applies to data practices involving users’ sensitive personal information, including biometric identifiers, precise geolocation, and financial account details. The sensitivity annotation aligns with several standards, including the EU General Data Protection Regulation (GDPR), the US National Institute of Standards and Technology (NIST) SP 800-122 guidelines,

and the Chinese standard GB/T 35273-2020, each representing regional and national regulations. For details, see Appendix C.

Rewritten Sentence provides user-friendly and comprehensible versions by rewriting and paraphrasing sentences labeled as “Importance”, “Sensitive”, and “Risk”.

To ensure that privacy policies reflect the latest trends and legal requirements, while also maintaining representativeness and diversity, we source the privacy policies of the top 100 most downloaded applications from two major app marketplaces: *Google Play* and the *App Store* in the English-speaking world. After excluding duplicates and instances where different applications share the same privacy policy, we obtain 139 distinct and representative privacy policies. These documents, current as of October 2023, cover a wide range of mainstream application types, including shopping, live streaming, sports, navigation, and various game genres such as real-time battles, sports competitions, simulation management, and board games. Given that subtle semantic differences in privacy policies can have legal implications, and that smaller models are still widely used in edge deployments, we employ regular expressions to segment these privacy policies into individual sentences for detailed annotation and interpretation.

Finally, we permit each segmented sentence to have multiple labels. The annotation work for this English privacy policy dataset is carried out by five experts, each holding a master’s degree with a legal background and possessing lawyer certification. They possess advanced English language skills, with proficiency surpassing the TOEFL and IELTS graduate admission requirements. Among these five annotators, three have previously participated in the Chinese privacy policy annotation of the CAPP-130 dataset (Zhu et al., 2023). Before starting this project, all annotators undergo systematic training covering annotation guidelines and typical examples to ensure annotation quality and stability. A preliminary analysis of five annotated documents yields a *Cohen’s Kappa* score of 0.892, indicating strong agreement. To account for the multi-label and multi-task nature of the dataset, this score was calculated by considering all classification categories, including Special Marking, Importance, Risk, and Sensitivity. We computed the agreement across these tasks to ensure the reliability of the integrated schema. This high level of

Topic	Num	Pct	Med.	Avg.
First	3667	11.8%	23.0	27.2
Permission	153	0.5%	22.0	23.2
Third	3019	9.8%	25.0	29.0
Usage	3159	10.2%	23.0	28.8
Retention	974	3.2%	25.0	28.2
Security	1069	3.5%	23.0	25.4
Specific	1858	6.0%	23.0	25.9
Control	3772	12.2%	22.0	24.8
Contact	1229	4.0%	19.0	21.0
Change	795	2.7%	18.0	20.2
Cease	13	0.04%	23.0	31.2
Important	15795	51.04%	24.0	27.3
Risk	577	1.9%	26.0	29.7
Sensitivity	374	1.21%	26.0	28.9

Table 1: Statistical Information of APPSI-139.

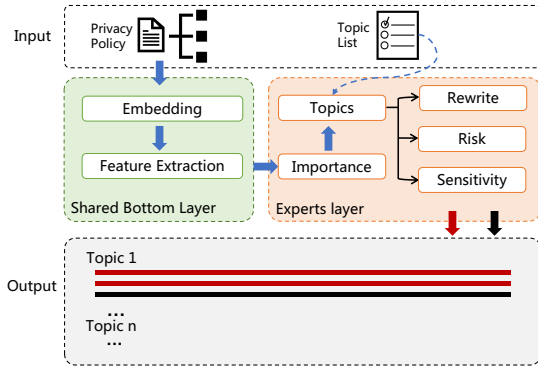


Figure 3: The framework of TCSI-pp-V2.

agreement informs our decision to adopt the strategy of assigning each expert to a single subtask, which includes the five specific tasks: identifying sentences related to “Importance”, “Risk”, “Sensitivity”, “Data Practice Categories”, or those requiring a “Rewritten”. During formal annotation, to transparently handle edge cases, annotators document ambiguous clauses. These are then resolved through discussion, with final adjudication by senior reviewers.

This results in a high-quality corpus named APPSI-139 (Application Privacy Policies Summarization and Interpretation, set of 139). This corpus comprises 36,351 annotations for 30,877 sentences, including 15,692 rewritten sentences.

The corpus enables analyze the composition of typical English privacy policies from a data practice perspective. Table 1 presents the statistics for APPSI-139. The Pct (*percentage*) represents the proportion of sentences in a specific data

practice category or with special markers relative to the total number of sentences. Since a sentence can have multiple labels, the total percentage may exceed 100%. *Mean* and *median* represent the average and middle values of sentence lengths, respectively. The average sentence length exceeds the median for all categories, indicating a right-skewed distribution. The APPSI-139 corpus shows an uneven distribution of data practice categories in English privacy policies. First-party collection (11.8%), third-party sharing (9.8%), editing/control (12.2%), and usage (10.2%) account for about 44% of the content, forming the core components. Rewritten sentences average 20 words, compared to 27 in the originals, reflecting a 26% reduction. In APPSI-139, the “Risk” label appears in 1.9% of sentences, highlighting the widespread privacy and data security risks faced by users of English-language applications.

3.2 The TCSI-pp-V2 Framework

To ensure the completeness and controllability of privacy policy summarization, we propose TCSI-pp-V2 (Topic-Controlled framework for Summarization and Interpretation of Privacy Policies), which represents a hybrid summarization approach based on end-to-end multi-task learning and builds upon the APPSI-139 corpus. As shown in Figure 3, the framework consists of four primary components:

Input: First, we preprocess any given English privacy policy $\forall P$ using *regular expressions* to extract a collection of sentences at the individual sentence level, denoted as $P = \{p_1, \dots, p_n\}$, where p_n refers to the n -th sentence in the policy. Second, users select the *topics* they wish to summarize from the *topics* that we provide.

Shared Bottom Layer: The Shared Bottom Layer processes the input signals and extracts key features for all tasks, reducing redundancy by learning shared representations. It begins with an embedding layer that captures the semantic content of the input text, represented as $P = \{p_1, \dots, p_n\}$. This layer converts the text into a numerical matrix, denoted as $E = \{e_1, \dots, e_n\}$. Next, the feature extraction function, $F_f(e_j, \theta_f)$, processes the embeddings E and extracts features, producing a new set, $Features = \{f_1, \dots, f_n\}$.

Experts Layer: Expert layers are components trained for specific tasks, with each expert layer specifically designed to handle different aspects of privacy policy summarization and interpretation.

Specifically, we train five distinct expert layers to accomplish the task of privacy policy summarization.

- **Importance identification** $F_i(f_j, \theta_i)$ takes all $\forall p$ corresponding features f_j as input and identifies the “Important” sentences. Here, θ_i represents a set of parameters obtained through training.
- **Topics identification** $F_t(f_j, \theta_t)$ indexes the features f_j of $\forall p_j | important = true$ to determine the data practice category (topic) to which these “important” sentences belong. Based on user-specified *topics*, we extract sentences consistent with them, producing the set *Filtered*.
- **Risk identification** $F_r(f_j, \theta_r)$ indexes the features f_j of $\forall p_j | p \in Filtered$ to identify sentences that potentially harm user data security.
- **Sensitivity identification** $F_s(f_j, \theta_s)$ indexes the features f_j of $\forall p_j | p \in Filtered$ to flag sentences related to sensitive personal information.
- **Text interpretation** $F_{rewrite}(f_j, \theta_{rewrite})$ rewrites important long sentences containing obscure professional and technical terms into simple and easily understandable short sentences. This addresses the issue of “incomprehensibility”. Specifically, $F_{rewrite}(f_j, \theta_{rewrite})$ uses an auto-regressive model $P(z_t | f_j; z_{1:t-1})$ to rephrase the original privacy policy into formulations that are easily comprehensible for the general public.

Output: The framework summarizes and interprets privacy policies based on preselected topics and highlights sentences that contain “Risk” and “Sensitivity” information to enhance their readability.

In summary, TCSI-pp-V2 introduces a shared encoder across all sub-tasks, replacing the five separate encoders in TCSI-pp, and incorporates task-specific logic optimization, thereby theoretically reducing encoding overhead by about 80%. Pseudocode appears in Appendix D.

4 Experiments

In this section, we introduced the experimental setup in Section 4.1, presented performance benchmarks for information extraction models based on APPSI-139 in Section 4.2, discussed performance benchmarks for the rewriting model in Section 4.3,

assessed TCSI-pp-V2’s readability through a social survey in Section 4.4, conducted a comprehensive evaluation of resource efficiency in Section 4.5, performed a comparative analysis with state-of-the-art Large Language Models in Section 4.6, and investigated the model’s robustness to varying input lengths in Section 4.7.

4.1 Base Models

We have implemented the TCSI-pp-V2 framework using mT5, Bert2GPT, XLNet2GPT, and Electra2GPT as base models. The model architecture has consisted of an encoder followed by four classification heads for the respective sub-tasks and a decoder for content rewriting. This design has allowed for the joint optimization of classification and generation tasks within a unified framework. For comparison, we also have implemented the TCSI-pp framework. For the four information identification sub-tasks, we have trained models based on the following Transformer encoders: RoBERTa (*xlm-roberta-base*)(Conneau et al., 2020), BERT (*bert-base-cased*)(Devlin et al., 2018), ERNIE (*ernie-2.0-base-en*)(Sun et al., 2020), PERT (*english-pert-large*)(Cui et al., 2022), DistilBERT (*distilbert-base-uncased*)(Sanh et al., 2019), ELECTRA (*electra-base-discriminator*)(Clark et al., 2020), XLNet (*xlnet-base-cased*)(Yang et al., 2019), DeBERTa (*deberta-v3-base*)(He et al., 2023), ALBERT (*albert-base-v2*)(Lan et al., 2019), and mT5 (*mT5-small*)(Xue et al., 2021). For the content rewriting sub-task, we have trained the TCSI-pp framework using Bert2Bert, Bert2GPT, ERNIE2GPT, and mT5. All of these base models have been available from the Hugging Face model repository HuggingFace¹. To ensure that the input sequence does not exceed the maximum position limit of the pre-trained models, we have truncated documents by removing the ending part. Data splitting details appear in Appendix E.

4.2 Information Extraction

Table 2 has presented the performance of models trained on the APPSI-139 corpus for information extraction. The table has included four multi-task learning TCSI-pp-V2 models and nine single-task learning TCSI-pp baseline models. The tasks “Importance,” “Risk,” and “Sensitivity” have been binary classification, while “Topic” has involved

¹<https://huggingface.co/>

TCSI-pp	Method	Topics		Important		Risk		Sensitive	
		Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
V1	RoBERTa	80.77	79.09	69.95	69.90	95.27	61.60	97.28	71.53
	Bert	81.00	79.63	76.87	76.21	95.54	61.58	97.07	62.51
	alBert	73.28	72.34	75.09	73.94	97.82	60.99	97.82	60.98
	Pert	79.82	78.58	78.84	78.20	96.74	49.17	97.93	49.48
	ERNIE	81.43	79.56	78.29	77.55	96.41	62.41	97.06	69.25
	DistilBert	80.55	79.18	77.73	77.03	95.70	58.18	97.34	60.43
	DeBERTa	78.47	76.33	79.35	78.34	99.27	66.55	97.01	68.33
	XINet	81.39	80.13	77.90	77.36	96.03	59.73	96.90	66.45
	Electra	81.91	80.03	78.00	77.22	96.08	60.69	96.90	62.20
	Mean	79.85	78.31	76.89	76.19	96.54	60.01	97.26	63.46
V2	mT5	78.18	77.12	73.93	73.55	95.60	58.77	96.96	67.40
	XINet2gpt	80.79	78.79	76.25	74.99	95.98	63.39	96.68	64.09
	Electra2gpt	80.00	78.83	76.19	75.79	95.49	59.31	93.36	64.53
	Bert2gpt	79.68	77.97	76.74	75.81	95.81	60.80	96.74	63.45
	Mean	79.66	78.17	75.78	75.04	95.72	60.57	95.93	64.86

Table 2: Evaluation metrics for the information identification models.

TCSI-pp	Model	ROUGE-1	ROUGE-2	ROUGE-L	BERT-score	Bart-score
V1	mT5	0.7199	0.6296	0.7072	0.947	-1.56
	Bert2gpt	0.6477	0.5411	0.6312	0.930	-2.22
	Bert2bert	0.6784	0.5980	0.6610	0.939	-2.14
	ERNIE2gpt	0.5805	0.3931	0.5459	0.926	-2.26
	Mean	0.6566	0.5405	0.6363	0.935	-2.05
V2	mT5	0.7051	0.6111	0.6903	0.943	-1.68
	Bert2gpt	0.6548	0.5456	0.6375	0.933	-2.03
	XINet2gpt	0.6751	0.5623	0.6598	0.939	-1.786
	Electra2gpt	0.6653	0.5476	0.6491	0.938	-1.89
	Mean	0.6752	0.5666	0.6592	0.938	-1.84

Table 3: Evaluation metrics for the rewrite expert models.

multi-class classification. Due to data sparsity, the categories “Cease Operation” and “Permission Acquisition” have been excluded. Evaluation metrics have included Micro-F1 and Macro-F1 scores.

As shown in Table 2, under the TCSI-pp framework, DeBERTa performs best on binary classification, while XLNet and ELECTRA excel in multi-class tasks. In TCSI-pp-V2, XLNet2GPT slightly outperforms the other three models. Across single-task models (BERT, XLNet, DeBERTa, ELECTRA) and their multi-task counterparts, performance differences remain within 0.02, indicating that TCSI-pp-V2’s multi-task models match the performance of TCSI-pp while reducing computational costs. Further details regarding Precision, Recall, and F1 scores for the multi-class “Topic” classification are provided in Appendix F.

4.3 Rewritten and Interpretations

To evaluate the performance of the interpretations model in TCSI-pp-V2, we have used rewritten sentences from APPSI-139 for fine-tuning. Four interpretations models have been implemented based on the TCSI-pp and TCSI-pp-V2 frameworks, respectively. Table 3 has displayed the evaluation metrics for information identification models, including *ROUGE* (Lin, 2004), *Bert-score* (Zhang et al., 2020), and *Bart-score* (Yuan et al., 2021). These metrics provide a multi-dimensional assessment: ROUGE measures lexical overlap, while Bert-score and Bart-score capture semantic consistency and generative quality. Together, they robustly reflect the model’s alignment with human expert judgments and legal intent beyond simple word-matching.

Model	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Percentage
TCSI-pp-V2	22	27	27	20	20	20	14	14	20	23	39.06%
Llama3-70b	18	6	13	5	15	11	21	21	14	10	25.28%
GPT-4o	5	15	6	25	12	11	13	13	14	16	24.52%
Kimi	8	5	7	3	6	11	5	5	5	4	11.13%

Table 4: Questionnaire statistics of readability.

Table 3 shows that mT5 outperforms other models in both the single-task setting based on TCSI-pp and the multi-task setting based on TCSI-pp-V2, with a slight advantage in single-task training. However, for Bert2GPT, the model trained with multi-task learning surpasses the single-task version across all metrics. Moreover, in the rewriting task, multi-task models consistently achieve higher average scores than single-task models, with particularly notable improvements in accuracy and recall. This suggests that multi-task training maintains or even enhances performance in generation tasks, as related tasks share common features and the model benefits from learning more generalized representations through joint training.

4.4 Readability Questionnaire Survey of TCSI-pp-V2

To evaluate the readability of the privacy policy summarization system based on TCSI-pp-V2, we have conducted a social survey on WJX². The survey has consisted of 10 single-choice questions, each presenting a paragraph from a privacy policy alongside summaries generated by four models: our model, CPT-4o³, Llama3-70b⁴, and Kimi (Moonshot V1)⁵. We have invited participants to select the summary they have found most readable, based on criteria such as comprehensibility, completeness, fidelity, and conciseness. In total, we have collected 53 valid responses, as shown in Table 4. The respondents, aged between 18 and 40, all have received university-level education. The summarization examples, instruction templates, and detailed statistics of the survey participants are provided in Appendix G.

According to Table 4, 39.06% of participants rate TCSI-pp-V2 summaries as the most readable based on comprehensibility, completeness, fi-

²<https://www.wjx.cn/>

³<https://openai.com/index/hello-gpt-4o/>

⁴<https://huggingface.co/meta-llama/Meta-Llama-3-70B>

⁵<https://kimi.moonshot.cn/>

Metric	TCSI-pp	TCSI-pp-V2
Encoding Time (s)	92.66	24.72
Task Inference Time (s)	99.28	98.54
Total Time (s)	191.94	123.26
Avg. Time per Sentence (s)	0.101	0.065
Inference Memory (MB)	7343	2766

Table 5: GPU Memory and Time Cost Comparison.

delity, and conciseness, followed by Llama3-70b (25.28%), GPT-4o (24.52%), and Kimi (11.13%). Additionally, TCSI-pp-V2 leads in seven out of ten questions, outperforming the other models overall. These results indicate that TCSI-pp-V2 demonstrates excellent overall readability when summarizing application privacy policies, outperforming Llama3-70b, GPT-4o, and other LLMs in privacy policy summarization and interpretation task.

4.5 Resource Efficiency

To evaluate the computational efficiency and practical deployability of our framework, we have measured the GPU memory usage and inference latency of TCSI-pp and TCSI-pp-V2 (both utilizing the mT5-small backbone). The evaluation has been conducted on a representative subset of 10 privacy policies comprising 1,893 sentences, with the metrics detailed in Table 5.

The results demonstrate that TCSI-pp-V2 achieves substantial efficiency gains. Specifically, the encoding time is reduced by approximately 73%, and the average processing time per sentence is nearly halved. Furthermore, the peak GPU memory overhead significantly drops from 7,343 MB to 2,766 MB. These improvements confirm that our optimized shared-encoder architecture effectively mitigates computational redundancy, making it more suitable for real-time applications without sacrificing inference accuracy.

4.6 Comparative Analysis with Large Language Models

To benchmark our framework against contemporary state-of-the-art baselines, we have compared TCSI-pp-V2 with several prominent Large Lan-

Task	Metric	Qwen3-8B	Llama3-8B	GPT-4o-mini	Gemini-2.5	TCSI-pp-V2
Topic	Micro F1	47.85	30.77	50.36	65.38	78.18
	Macro F1	44.44	24.65	32.83	6.42	77.12
Important	Micro F1	63.50	52.53	51.00	73.33	73.93
	Macro F1	62.11	52.45	49.36	71.29	73.55
Risk	Micro F1	85.53	96.00	89.34	85.05	95.60
	Macro F1	49.30	48.98	51.24	45.95	58.77
Sensitive	Micro F1	45.37	28.01	38.64	23.33	96.96
	Macro F1	32.52	21.88	27.87	21.76	67.40
Rewritten	ROUGE-L	0.4541	0.4156	0.4286	0.4776	0.6903
	BERTScore	0.8970	0.8520	0.8950	0.9070	0.9430
	BARTScore	-2.76	-3.03	-2.89	-2.78	-1.68

Table 6: Performance Comparison between LLMs and TCSI-pp-V2 on the APPSI-139 Dataset.

guage Models (LLMs), including open-weight models (Qwen3-8B, Llama-3.1-8B) and proprietary systems (GPT-4o-mini, Gemini-2.5). These models have been evaluated on the APPSI-139 dataset using meticulously designed prompt engineering to handle both classification and abstractive rewriting tasks, with results shown in Table 6.

As shown in Table 6, TCSI-pp-V2 consistently outperforms the selected LLMs across most tasks and does so with a significantly smaller parameter count. This indicates that even though general-purpose LLMs possess vast general knowledge, they still struggle with the domain-specific nuances of privacy legalities, particularly in Topic and Sensitive classification. Moreover, our model maintains superior generation quality with higher ROUGE-L and BERTScore values. These results validate that specialized fine-tuning on expert-annotated data remains more effective than zero-shot prompting for high-precision legal analysis.

4.7 Robustness to Input Length

To assess the robustness of TCSI-pp-V2, we have conducted a comprehensive analysis focusing on extreme input distributions. Specifically, TCSI-pp-V2 has been evaluated on the 100 longest and shortest test samples across all classification and rewriting tasks. This evaluation has highlighted the framework’s adaptability to varying information densities, with detailed results in Table 7.

The experimental results demonstrate that TCSI-pp-V2 maintains high stability across the input length spectrum. We observe only minor performance fluctuations: exceptionally long policies exhibit a slight decrease in Risk classification accuracy due to their complex clause structures,

Task	Metric	Longest	Shortest	All
Topic	Micro F1	79.41	76.23	78.18
	Macro F1	77.20	75.09	77.12
Important	Micro F1	74.56	73.64	73.93
	Macro F1	71.85	72.35	73.55
Risk	Micro F1	94.74	94.34	95.60
	Macro F1	59.39	56.91	58.77
Sensitive	Micro F1	96.83	96.18	96.96
	Macro F1	66.92	66.76	67.40
Rewritten	ROUGE-L	0.6979	0.6542	0.6903
	BERTScore	0.9450	0.9430	0.9430
	BARTScore	-1.63	-1.75	-1.68

Table 7: Performance of Different Lengths.

while extremely short texts yield marginally lower ROUGE-L scores due to the scarcity of contextual cues. Nevertheless, the consistent performance across all subsets underscores the effectiveness of our multi-task architecture in processing diverse and structurally complex privacy documents.

5 Conclusion

To address the challenges of understanding complex and legally dense privacy policies, we introduce APPSI-139, the first parallel corpus of English privacy policies annotated by legal experts, aimed at providing user-friendly interpretations. We also propose the TCSI-pp-V2 framework, a hybrid summarization method based on multi-task learning that effectively balances computational efficiency and accuracy. Based on this framework, we develop a privacy policy summarization system, which, according to evaluations, outperforms general-purpose models like GPT-4o in readability and reliability. This system helps users make more informed and rational privacy decisions, enhancing the comprehension and interpretation of English privacy policies.

Limitation

Currently limited to English-language privacy policies, we aim to expand our framework to cover additional jurisdictions, multilingual settings and other privacy document types in the future. Additionally, as the system relies on model inference and training data, it may, in some cases, overlook key information or generate different interpretations, which could impact the completeness and accuracy of the summary. This is a common risk associated with all generative models when summarizing privacy policies.

Acknowledgement

The present research was supported by the National Key Research and Development Program of China (Grant No. 2023YFE0116400), NSFC (U23A20306, 62472378, U25A20430), the Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars (LR25F020001), and the Zhejiang Province Pioneer Plan (2024C01074, 2025C01084). We would like to thank the anonymous reviewers for their insightful comments.

References

- Wasi Uddin Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. [Policyqa: A reading comprehension dataset for privacy policies](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, Findings of ACL, pages 743–749. Association for Computational Linguistics.
- Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.
- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.
- Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, and 1 others. 2020. Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text. In *Proceedings of The Web Conference 2020*, pages 1943–1954.
- Duc Bui, Kang G Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated extraction and presentation of data practices in privacy policies. *Proc. Priv. Enhancing Technol.*, 2021(2):88–110.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [ELECTRA: Pre-training text encoders as discriminators rather than generators](#). In *8th International Conference on Learning Representations, Addis Ababa, Ethiopia, April 26-30, 2020*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.
- Yiming Cui, Ziqing Yang, and Ting Liu. 2022. [PERT: Pre-training BERT with permuted language model](#). *CoRR*, abs/2203.06906.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Tatiana Ermakova, Annika Baumann, Benjamin Fabian, and Hanna Krasnova. 2014. Privacy policies and users trust: Does readability matter? In *Americas Conference on Information Systems*.
- Yu Fang, Xiaoqi Pang, Qiyang Yu, Fan Min, Xuemei Cao, Pan Tao, and Tianrui Li. 2026. [Alignment in large vision language models: A survey](#). *Inf. Fusion*, 133:104294.
- Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. [Sentence simplification via large language models](#). *CoRR*, abs/2302.11957.
- Yu Fu, Deyi Xiong, and Yue Dong. 2023. [Inverse reinforcement learning for text summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6559–6570, Singapore. Association for Computational Linguistics.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.
- Armin Gerl, Nadia Bennani, Harald Kosch, and Lionel Brunie. 2018. [LPL, towards a gdpr-compliant privacy language: Formal definition and usage](#). *Trans. Large Scale Data Knowl. Centered Syst.*, 37:41–80.
- Elias Grünewald and Frank Pallas. 2021. Tilt: A gdpr-aligned transparency information language and toolkit for practical privacy engineering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 636–646.
- Hamza Harkous, Kassem Fawaz, Rémi Lebre, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. [Poli-sis: Automated analysis and presentation of privacy](#)

- policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548, Baltimore, MD. USENIX Association.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Nakayama Hiroki, Kubo Takahiro, Kamura Junya, Taniguchi Yasufumi, and Liang Xu. 2018. [doccano: Text annotation tool for human](#). Software available from <https://github.com/doccano/doccano>.
- Ziyuan Huang, Jiaming Tang, Manish Karir, Mingyan Liu, and Armin Sarabi. 2024. Analyzing corporate privacy policies using ai chatbots. In *Proceedings of the 2024 ACM on Internet Measurement Conference*, pages 505–515.
- Moniba Keymanesh. 2022. *Adaptive Summarization for Low-resource Domains and Algorithmic Fairness*. Ph.D. thesis, The Ohio State University.
- Moniba Keymanesh, Micha Elsner, and Srinivasan Sarthasrathy. 2020. Toward domain-guided controllable summarization of privacy policies. In *NLLP@KDD*, pages 18–24.
- Konrad Kollnig, Anastasia Shuba, Max Van Kleek, Reuben Binns, and Nigel Shadbolt. 2022. [Goodbye tracking? impact of ios app tracking transparency and privacy labels](#). *CoRR*, abs/2204.03556.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. [ALBERT: A lite BERT for self-supervised learning of language representations](#). *CoRR*, abs/1909.11942.
- 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 *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 588:1–588:24. ACM.
- Zhigen Li, Jianxiang Peng, Yanmeng Wang, Tianhao Shen, Minghui Zhang, Linxi Su, Shang Wu, Yihang Wu, Yuqian Wang, Ye Wang, Wei Hu, Jianfeng Li, Shaojun Wang, Jing Xiao, and Deyi Xiong. 2024. [Planning with large language models for conversational agents](#). *CoRR*, abs/2407.03884.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. 2018. Towards automatic classification of privacy policy text. *School of Computer Science Carnegie Mellon University*.
- Razieh Nokhbeh Zaeem, Ahmad Ahabab, Josh Bestor, Hussam H Djadi, Sunny Kharel, Victor Lai, Nick Wang, and K Suzanne Barber. 2022. [Privacy-check v3: Empowering users with higher-level understanding of privacy policies](#). In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1593–1596.
- Razieh Nokhbeh Zaeem, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava, and K Suzanne Barber. 2020. [Privacy-check v2: A tool that recaps privacy policies for you](#). In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3441–3444.
- Jonathan A Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. [Is chatgpt a general-purpose natural language processing task solver?](#) In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 1339–1384.
- Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958.
- Shivanshu Saini, Sanskriti Pathak, Suruchi Singh, Bharat Bhardwaj, and Saransh Sharma. 2025. Text summarization with llm: A comparison of transformer and non-transformer models. In *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)*, volume 3, pages 23–28. IEEE.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter](#). *CoRR*, abs/1910.01108.
- 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*, pages 2774–2779.

- Dan Shi, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang Liu, Xinwei Wu, Zishan Guo, Linhao Yu, Ling Shi, Bojian Jiang, and Deyi Xiong. 2024. [Large language model safety: A holistic survey](#). *CoRR*, abs/2412.17686.
- Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.
- Ravi Inder Singh, Manasa Sumeeth, and James Miller. 2011. A user-centric evaluation of the readability of privacy policies in popular web sites. *Information Systems Frontiers*, 13:501–514.
- Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. [ERNIE 2.0: A continual pre-training framework for language understanding](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8968–8975. AAAI Press.
- Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, and Lei Fan. 2023. [Policygpt: Automated analysis of privacy policies with large language models](#). *CoRR*, abs/2309.10238.
- Noriko Tomuro, Steven Lytinen, and Kurt Hornsberg. 2016. Automatic summarization of privacy policies using ensemble learning. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, pages 133–135.
- Shuai Wang, Xiang Zhao, Bo Li, Bin Ge, and Daquan Tang. 2017. Integrating extractive and abstractive models for long text summarization. In *2017 IEEE International Congress on Big Data (BigData Congress)*, pages 305–312. IEEE.
- 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, and 1 others. 2016. 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.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. [XLNet: Generalized autoregressive pretraining for language understanding](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [Bartscore: Evaluating generated text as text generation](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.
- Sara Zayed, Mostafa Ezzat, and Hesham A Hefny. 2025. Automatic text summarization: A review of approaches, challenges, and future directions. *Journal of Computer Science & Technology*, 25.
- Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2025. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Kaifa Zhao, Le Yu, Shiyao Zhou, Jing Li, Xiapu Luo, Aemon Yat Fei Chiu, and Yutong Liu. 2022. [A fine-grained chinese software privacy policy dataset for sequence labeling and regulation compliant identification](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10266–10277. Association for Computational Linguistics.
- Hou Zhu, Mingxin Zhang, and Yonghe Lu. 2018. An empirical study on privacy policy reading intention of social media users. *Journal of the China Society for Scientific and Technical Information*, 37(4):362–371.
- pengyun Zhu, Long Wen, Jinfei Liu, Feng Xue, Jian Lou, Zhibo Wang, and Kui Ren. 2023. [Capp-130: A corpus of chinese application privacy policy summarization and interpretation](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46773–46785. Curran Associates, Inc.
- Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. 2024. [Multilingual large language models: A systematic survey](#). *CoRR*, abs/2411.11072.
- Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66.

A Key Information of APPSI-139

Intended Users APPSI-139 is designed for researchers and practitioners in the fields of natural language processing, privacy policy analysis, and AI systems. It supports a wide range of applications, including multi-task learning, text summarization, and text classification. The dataset is also suitable for policymakers and legal scholars interested in examining how privacy practices are communicated in consumer-facing documents.

Hosting and Maintenance Plan All datasets in APPSI-139 are hosted and version-tracked via GitHub and are publicly available for direct download. Our core developing team is committed and has the resources to maintain and actively develop APPSI-139 for, at minimum, the next three years. We plan to grow APPSI-139 in several dimensions by including new learning datasets and leaderboards. We welcome external contributors.

Computing resources We use a server with an NVIDIA V100 GPU, Intel(R) Xeon(R) CPU with 128GB RAM for all empirical experiments in this manuscript.

Code availability The source code, annotation guidelines, and dataset are provided in a public repository, which can be accessed via <https://github.com/EnlightenedAI/APPSI-139>.

Ethics Statement The development and dissemination of the APPSI-139 dataset adhere to stringent ethical standards to ensure the integrity of the data, and the responsible use of the information. This transparency ensures that users of the APPSI-139 dataset understand the origin of the data and the context in which it was collected.

Transferability and Scalability The APPSI-139 dataset is designed with high transferability across different legal regimes and domains. Our annotation schema is hierarchical and fine-grained, abstracting common regulatory elements such as data collection, sharing, and sensitive data categories that are consistent across jurisdictions. For regulatory requirements in other regions, this schema is locally extendable or fine-tuneable without rebuilding the entire framework. In future work, our focus remains on exploring these localized extensions to support multilingual applications and cross-jurisdictional use.

Intended uses and dataset documentation

The dataset is designed for automatic summarization and rewriting of privacy policies, aiming to help users make more informed and rational privacy decisions. By providing high-quality privacy policy summaries, users can more easily understand complex legal terms and privacy agreements, enabling them to make wiser choices when handling personal data and privacy rights. The dataset supports research and implementation across various methods, including natural language classification and retrieval, text generation, in-context learning, and instruction fine-tuning. In addition to labeling and retrieving privacy policy content, it can also optimize LLMs through in-context learning and instruction fine-tuning, adjusting summary content based on user needs, thereby enhancing the model's performance in privacy policy applications.

Detailed documentation following the “Datasheets for Datasets” guidelines (Geburu et al., 2021) is provided in Section H.

Accessibility

1. Links to access the dataset and its metadata <https://github.com/EnlightenedAI/APPSI-139>.
2. The data is saved in a JSON format, where an example is shown in the README.md.
3. Research group will maintain this dataset on the official Github account.
4. CC BY 4.0. (<https://creativecommons.org/licenses/by/4.0/>)

B Data Practice Category

Data Practice Category Information also known as *Topic*, is used to describe the category of the sentence or term in privacy policies. It includes:

- **First Party Collection:** The types of user information collected by the service provider, the purpose of collection, and whether providing such information is mandatory.
- **Permission Acquisition:** How the service provider obtains application permissions from users, the purpose of these permissions, and whether they are mandatory.

Category	Example Items
Personal Property Information	Bank account, authentication information (password), deposit information (fund amounts, payment/receipt records), property information, credit records, credit reports, transaction and consumption records, transaction history, virtual currency, virtual transactions, game exchange codes, etc.
Personal Health Information	Illnesses, hospitalization records, medical prescriptions, laboratory reports, surgery/anesthesia records, nursing records, medication records, drug/food allergy information, reproductive information, past medical history, family medical history, infectious disease history, etc.
Biometric Information	Genetic data, fingerprints, voiceprints, palm prints, ear shapes, iris, facial recognition features, etc.
Personal Identity Information	ID card, military ID, passport, driver's license, work permit, social security card, residence permit, etc.
Other Information	Sexual orientation, marital history, religious beliefs, unpublished criminal records, communication records and content, contact lists, friend lists, group lists, travel history, web browsing history, accommodation information, precise location data, etc.

Table 8: Examples of personal sensitive information.

- **Third Party Sharing:** The purposes, types, and methods of sharing or disclosing user information to third parties.
- **Usage:** How user data is used, including for data analysis, personalized recommendations, etc.
- **Data Retention:** How user information is stored, including the duration and location of storage.
- **Data Security:** How user information is protected and measures for handling data breaches.
- **Edit/Control:** How users manage and handle their provided personal data.
- **Specific Audiences:** Describes privacy practices specific to particular user groups, such as minors or residents of specific regions.
- **Contact Information:** How to contact the service provider.
- **Policy Change:** How users are notified about changes to the privacy policy.
- **Cease Operation:** How user data is handled when the service ceases operations.

C Personal Sensitive Information

Personal sensitive information refers to data that, if disclosed, misused, or disclosed without authorization, may pose a threat to an individual's safety and property. It can also lead to damage to personal reputation, physical and mental health, or result in discriminatory treatment. Generally, personal information related to children under 14 years old (inclusive) and any information directly linked to an individual's privacy is considered sensitive, as outlined in the Chinese national standard *GB/T 35273-2020*. The following criteria can help determine whether information qualifies as personal sensitive information:

1. **Disclosure:** When personal information is disclosed, the individual and the organization or institution collecting or processing it lose control over its distribution, resulting in uncontrolled spreading and usage. In some cases, personal information, once leaked, may be used against the individual's will or in conjunction with other data, posing a significant risk to the person's rights. This type of information should be categorized as personal sensitive information. For example, if someone's ID card is used to register a phone number or open a bank account without their permission, it should be considered personal

Algorithm 1 TCSI-pp-V2 framework.

Input: Privacy policy P ; Specified $topics \in Topics$.
Output: Summarization P_{ats} .
Initialize: $P = \{p_1, \dots, p_n\} \leftarrow Preprocessing(P)$; $Filtered = list()$
#Step 1: Five trained *experts* carry out auto summarization.
For $\forall p_j \in P$:
 $f_j \leftarrow F_f(F_e(p_j, \theta_e), \theta_f)$
 If $F_i(f_j, \theta_i)$ is *True*: $topic = F_t(f_j, \theta_t)$
 If $\exists topic$ in $topics$:
 $Filtered \leftarrow topic, r_j = F_r(f_j, \theta_r), s_j = F_s(f_j, \theta_s), rewrite_j = F_{rewrite}(f_j, \theta_{rewrite})$ cc
#Step 2: Examine into a more understandable Summarization.
For $\forall topic \in topics$: $P_{ats} \leftarrow title_topic$
 #Write the title of $topic$ to P_{ats} .
 For $\forall rewrite_i \in topic$ in $Filtered$:
 If r_i or s_i is *True*: $P_{ats} \leftarrow rewrite_i$ with highlight
 Else: $P_{ats} \leftarrow rewrite_i$
Return P_{ats}

	all samples	training samples	validation samples	testing samples
Important	30877	24710	3088	3088
Risk	15579	12607	1576	1576
sensitive	15579	12607	1576	1576
Topic	15579	12607	1576	1576
Rewritten	15677	12543	1567	1567

Table 9: The statistics of train-validation-test data.

sensitive information.

- Illegal Provision:** Certain personal information becomes a significant risk to the individuals rights when shared without consent, especially if its spread beyond the intended scope. Such information should be regarded as personal sensitive information. For instance, sexual orientation, banking details, and medical history related to infectious diseases should be considered sensitive if disclosed without consent.
- Abuse:** Some personal information, when used beyond its authorized limits or for purposes other than originally intended, may pose a substantial risk to an individual’s rights. This information should be classified as personal sensitive information. For example, if an insurance company uses a persons health information to determine premiums for marketing purposes without obtaining the individual’s consent, this is an abuse of personal sensitive information.

We provides examples of personal sensitive information in Table 8.

D TCSI-pp-V2 framework

The pseudocode of the TCSI-pp-V2 framework is shown in Algorithm 1.

E Data Splitting

We split the APPSI-139 corpus into training data, validation data, and testing data at a ratio of 80:10:10. This approach effectively prevents potential data leakage and ensures the independence of the testing units. The details of each corpus are shown in Table 9. The source code, annotation guidelines, and dataset are available in a public repository at <https://github.com/EnlightenedAI/APPSI-139>.

F Detailed Results for Topic Classification

Table 10 has provided a comprehensive evaluation of Precision, Recall, and F1 scores for multi-class classification on Topic, covering two mod-

Topic	TCSI-pp						TCSI-pp-v2					
	Xlnet			Electra			Xlnet2gpt			Electra2gpt		
	P	R	F	P	R	F	P	R	F	P	R	F
First	81.5	82.6	82.1	88.2	79.4	83.5	84.4	80.3	82.3	90.2	71.6	79.9
Third	81.3	78.1	79.7	81.3	78.1	79.7	79.3	82.4	80.8	81.0	71.8	76.1
Usage	85.9	78.9	82.2	88.2	80.6	84.3	88.7	75.9	81.9	83.3	80.6	82.0
Retention	81.6	85.7	87.9	79.5	79.5	79.5	80.0	72.7	76.2	81.5	75.0	78.1
Security	83.8	76.0	79.7	70.4	74.2	72.3	73.2	79.6	76.3	79.7	67.7	73.3
Specific	90.3	85.7	87.9	89.5	89.3	89.4	91.2	83.9	87.4	83.0	91.6	87.1
Control	83.8	76.0	79.7	78.7	78.7	78.7	89.5	68.0	77.3	81.7	77.3	79.5
Contact	87.1	80.6	83.7	91.5	80.6	85.7	89.8	79.1	84.1	94.6	79.1	86.2
Change	75.8	65.3	70.1	75.7	60.4	67.2	86.6	49.3	62.8	73.2	62.5	67.4

Table 10: Evaluation metrics for the Topics identification models.

Characteristic	Attribute	Subtotal	Percentage
Gender	Male	27	50.94%
	Female	26	49.06%
Age	< 18	0	0%
	18-25	38	71.70%
	26-30	12	22.64%
	31-40	3	5.66%
	>40	0	0%
Education	Associate & below	0	0%
	Bachelor’s	23	43.39%
	Master’s & above	30	56.60%

Table 11: Basic statistical information of the survey participants.

els with the same encoder modules (XLNet and ELECTRA) under the TCSI-pp and TCSI-pp-V2 frameworks.

According to Table 10, XLNet and ELECTRA achieve comparable performance across topics under both the TCSI-pp and TCSI-pp-V2 frameworks. In terms of F1 scores, XLNet performs best overall, with scores above 0.79 for all topics except Policy Change. Across all metrics, models based on TCSI-pp and TCSI-pp-V2 show mixed results, with no significant differences, further indicating that TCSI-pp-V2’s multi-task models match the performance of TCSI-pp while offering an efficiency advantage. Additionally, all models score relatively low on Policy Change, likely due to its small dataset proportion.

G Additional of Readability Survey

G.1 Basic statistical information of the survey participants

Table 11 presents the basic statistical information of the survey participants in the privacy policy

readability survey.

G.2 Examples of Summarization

To objectively assess the performance of TCSI-pp-V2, fine-tuned based on the mt5-small model, alongside large language models GPT-4o, Llama3-70b, and Kimi (Moonshot V1) in the task of privacy policy summary generation, we anonymized a privacy policy document and selected a portion of its content for testing, which is displayed in main paper Figure 4. We then presented the summary results of TCSI-pp-V2, GPT-4o, Llama3-70b, and Kimi (Moonshot V1), shown in Figures 5, 6, 7, and 8, respectively. The results indicate that TCSI-pp-V2 has a distinct advantage in generating structured summaries, capturing the essence of the original text with conciseness and precision. However, the summaries produced by other general language models appear more fragmented and lack coherence, as demonstrated by Llama3-70b and Kimi; or they exhibit shortcomings in identifying key information and effectively synthesizing it, such as GPT-4o, which almost verbatim repro-

Privacy Policy

We value your ("User," "You," or "Your") privacy and recognize the sensitivity of your personal information. We are committed to protecting your personal information and using it only as appropriate to provide you with the best possible service, products, and opportunities, as described in this Privacy Policy. This Privacy Policy explains the online information practices of GGroup Limited (collectively, "Company," "we," "us," or "our") which hosts the website of http://www.***.com/ and the mobile application PT (collectively, the "Application"). Please note that when using the Application, this Privacy Policy should be read in conjunction with our Terms of Use. By accessing the Application, you acknowledge that the Privacy Policy and the Terms of Use govern your use of the Application.

Privacy and personal data protection principles vary from one country to another. When you access or visit a third party website or application, you are subject to the privacy policy of that website or application. Please read the third party's privacy policy to determine the privacy practices that apply to information or data maintained by the website or application.

I. What Information We Collect and How It Is Used

If you collaborate on or become a member of an Organization, then its Account owners may receive your personal data, for example by having the ability to view your activity in the Organization's access log.

We may aggregate or deidentify your information and may use, share, rent or sell aggregated or deidentified information for any purpose, and such information is not subject to this Privacy Policy.

For contractors and vendor representatives, we may process identification information and work-related information, as necessary to manage our relationship with you and your employer, which is necessary for the performance of the services agreement, and to establish, exercise or defend potential legal claims.

We also use personal data to manage our relationships and contracts with account owners and others, including billing, compliance with contractual obligations, facilitating payment to third-party developers in relation to purchases made through the App Marketplace, and related administration.

We may disclose aggregated information about our users, and information that does not identify any individual or device, without restriction.

In our services, we may need to collect and process your personal sensitive information, including but not limited to race, political opinions, religious beliefs, health conditions, or sexual orientation.

Our services allow users to generate content, such as comments or uploaded images. Please note that this content may be seen or shared by other users, and we are not responsible for the security or privacy of user-generated content.

We take appropriate technical and organizational measures to protect your personal data from unauthorized access and misuse. We only retain your data as necessary and within the limits permitted by law.

We may collect the content you create, upload, generate, or access through the Services, including the photographs, images, and videos you upload, video templates you generate or view, comments you make, and other actions you take on the Services, such as liking content, including associated metadata (such as when, where, and by whom the content was created) ("User Content").

Figure 4: Application Privacy Policy.

duces the entire privacy policy document.

G.3 Instruction Template and Evaluation Setup for LLMs.

To ensure fairness and consistency, we have provided structured guidance to the large language models (LLMs) via the prompt template detailed in Algorithm 2. This template includes comprehensive category definitions to establish clear task boundaries and explicit requirements for rewriting and formatting. In our comparative experiments, we utilize this standardized prompt with specific hyperparameter settings (e.g., temperature = 0.7 and top-p = 0.9) to ensure reproducible generation across models. These measures were implemented to minimize ambiguity and ensure that all models are evaluated under the same well-defined conditions.

G.4 Evaluation Methodology

To capture the overall preferences of human evaluators, we evaluate the summaries across four critical dimensions: *comprehensibility* (how easy the summary is to understand), *completeness* (whether the summary covers all key aspects without omissions), *fidelity* (the accuracy and consistency of the summary with the original text, ensuring that no misunderstandings or distortions occur), and *con-*

ciseness (the ability to convey the essential points without unnecessary elaboration). For the readability survey, these dimensions are combined into a single composite score. While we acknowledge that this approach may obscure some fine-grained differences, the composite score effectively captures the trade-offs participants make across these dimensions and their holistic judgment of summary quality.

To ensure a rigorous and unbiased comparison, all summaries are normalized to maintain consistent formatting across different systems. During the survey, the summaries are presented to each evaluator in a randomized order to mitigate potential order biases. Furthermore, all materials are presented in a consistent monochrome format, aligned with the baseline methods. The colored highlights and nodes in Figure 5 are provided solely as a visual aid for the readers and were absent from the actual human evaluation process.

A critical challenge in legal and privacy policy summarization is the trade-off between readability and factuality. While simplifying complex legal jargon enhances user comprehension, excessive abstraction may risk omitting nuanced legal obligations or distorting factual accuracy. Our framework addresses this by leveraging expert-annotated interpretation data, ensuring that the

Summarization by TCSI-pp-V2

1. How We Collect Your Information

Sensitive Warning: (1) In our services, we may need to collect and process your sensitive personal information.

2. How We Share Your Information

Risk Warning: (1) Our services enable user-generated content like comments and uploads, but we're not responsible for its security or privacy.

3. How We Use Your Information

- (1) The account owners may receive your personal data, for example by having the ability to view your activity in the Organization's access log.
- (2) We may aggregate or de-identified information for any purpose, and such information is not subject to this Privacy Policy.
- (3) We may process identification information and work-related information for contractors and vendor representatives
- (4) We also use personal data to manage our relationships and contracts with account owners and others.
- (5) We may disclose aggregated information about our users, and information that does not identify any individual or device, without restriction.
- (6) We take appropriate technical and organizational measures to protect your personal data from unauthorized access and misuse.
- (7) We may collect the content you create, upload, generate, or access through the Services.

4. How You Can Manage Your Information

Nothing was found related to "How You Can Manage Your Information".

Note:

Risk Warning: Indicates potential security or privacy risks associated with user data.

Sensitive Warning: Indicates sensitive personal information being collected or processed.

Figure 5: Summarization by TCSI-pp-V2.

simplified output remains grounded in the original legal intent. Our human evaluation criteriaspecifically fidelity and completenessare designed to monitor this trade-off, ensuring that improved readability does not come at the expense of factual reliability.

H Datasheets for Datasets

H.1 Motivation

- **For what purpose was the dataset created?** (Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)

To address the common issues of “lengthiness” and “incomprehensibility” that the public often encounters when reading privacy policies, and to prevent information security risks caused by excessive authorization, we have carried out innovative work. Although existing English privacy policy corpora can alleviate the issue of “lengthiness” to some extent, they lack practical solutions for the “incomprehensibility” problem caused by professional jargon, technical terms, and complex sentence structures.

Therefore, we have released the APPSI-139 corpus, which consists of 139 English application privacy policies and 30,877 sentences. In this corpus, we have annotated 11 types of data practices and three special markers (to identify the *Importance*, *Risk*, and *Sensitivity* of privacy policies), and ultimately rewritten the important sentences through Text interpretation, i.e., reformulated them into a language form that is easier to understand. This is different from other English privacy policy summary datasets, as APPSI-139 provides rewritten content, contributing to the development of generative automatic summarization technology in the field of privacy policies. Based on APPSI-139, we have conducted benchmark tests on the performance of classic deep learning models in information extraction and text generation tasks.

- **Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

DIVER, an interdisciplinary research group at Zhejiang University.

Summarization by GPT-4o

We value your privacy and understand the importance of your personal information. We are dedicated to protecting your information and using it appropriately to offer you the best service and products, as described in this Privacy Policy. This document explains how GGGroup Limited ("Company," "we," "us," or "our") handles your information when you use our website at http://www.***.com/ and the mobile app PT (collectively, the "Application"). By accessing the Application, you agree that this Privacy Policy and our Terms of Use apply to your use of the Application.

Privacy laws differ across countries. When you visit third-party websites or applications, their privacy policies apply. Always read their privacy policies to understand how they manage your data.

I. Information We Collect and Its Use

If you join an Organization, its Account owners can see your personal data, such as your activity in the Organization's access log.

Risk Notice: We may aggregate or de-identify your information for any purpose, and such information is not covered by this Privacy Policy.

For contractors and vendor representatives, we process identification and work-related information to manage our relationship with you and your employer, necessary for our services agreement, and to handle potential legal claims.

We also use personal data to manage relationships and contracts with account owners and others, including billing, compliance with contracts, facilitating payments to third-party developers, and related administration.

Risk Notice: We may disclose aggregated information about users and non-identifying information without restriction.

Sensitive Data: In our services, we may need to collect and process sensitive personal information, including race, political opinions, religious beliefs, health conditions, or sexual orientation.

Our services allow users to create content, such as comments or uploaded images. This content may be visible or shared with other users, and we are not responsible for its security or privacy.

We implement appropriate technical and organizational measures to protect your data from unauthorized access and misuse. We only keep your data as necessary and as permitted by law.

We may collect content you create, upload, or access through our services, including photos, images, videos, templates, comments, and other actions like liking content, along with associated metadata (such as when, where, and by whom the content was created) ("User Content").

Figure 6: Summarization by GPT-4o.

H.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)** (Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.)

The sentences in the privacy policy constitute instances of the APPSI-139, which include "Original Sentence", "Data Practice Category", "Special Marking", and the "Rewritten Sentence".

- **How many instances are there in total** (of each type, if appropriate)?

The dataset contains 139 English privacy policies of mainstream applications, including 30877 preprocessed sentences.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** (If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances,

because instances were withheld or unavailable).)

This dataset is a sample from a larger set. Details could be checked in Section 3.1.

- **What data does each instance consist of?** ("Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.)

Each instance has "Original Sentence", "Sentence ID" and annotations, including "Data Practices Category", "Special Marking", and "Rewritten Sentence".

- **Is there a label or target associated with each instance?** If so, please provide a description.

The privacy policy annotations are structured into three distinct components: "Data Practices Category", "Special Marking", and "Rewritten Sentence". For comprehensive details, refer to Section 3.1 of the main paper.

- **Is any information missing from individual instances?** (If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.)

Summarization by Llama3-70b

We value your privacy and are committed to protecting your personal information. This Privacy Policy explains how we collect, use, and share your information when you use our Application.

What Information We Collect and How It Is Used

We may collect different types of information from you, including:

- Personal data, such as identification information and work-related information, to manage our relationships and contracts with account owners and others.
- User-generated content, such as comments, uploaded images, and videos, which may be seen or shared by other users.
- Aggregated or deidentified information, which we may use, share, rent, or sell for any purpose.

Important Notes

Possible Risks

Third-Party Websites and Applications: When you access a third-party website or application, you are subject to their privacy policy, which may be different from ours.

User-Generated Content: We are not responsible for the security or privacy of user-generated content, which may be seen or shared by other users.

Sensitive Personal Information

Warning: We may need to collect and process sensitive personal information, including race, political opinions, religious beliefs, health conditions, or sexual orientation. We take appropriate measures to protect this information, but please be aware of the potential risks.

How We Protect Your Information

We take appropriate technical and organizational measures to protect your personal data from unauthorized access and misuse. We only retain your data as necessary and within the limits permitted by law.

Account Owners and Contractors

If you collaborate on or become a member of an Organization, the Account owners may receive your personal data, such as activity logs.

For contractors and vendor representatives, we may process identification information and work-related information to

Figure 7: Summarization by Llama3-70b.

Sentences that lack special marking do not necessitate rewriting. As these sentences pertain to mundane declarative documents, summarization and rewriting are deemed unnecessary.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** (If so, please describe how these relationships are made explicit.)

In affirmation, each instance is endowed with a distinctive identifier "id", which facilitates the correlation to its respective privacy policy.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** (If so, please provide a description of these splits, explaining the rationale behind them.)

Yes. We did a train-validation-test split on the dataset, see Appendix E.

- **Are there any errors, sources of noise, or redundancies in the dataset?** (If so, please provide a description.)

The creation of the APPSI-139 corpus involved annotations by legal experts, which

may introduce bias based on their individual interpretations and perspectives. While efforts have been made to ensure consistency, such as strict annotation standards, variations in interpretations may impact the quality of the corpus.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**(If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)

The APPSI-139 dataset under consideration is autonomous and does not necessitate or depend on external resources.

- **Does the dataset contain data that might be considered confidential (e.g., data that**

Summarization by Kimi

Privacy Policy Summarization

Who We Are: GGGroup Limited is the company behind the website and mobile app PT.

What This Policy Covers: This policy explains how we handle your personal information when you use our website or app.

Third-Party Sites: Be aware that other websites or apps have their own privacy policies.

What Information We Collect and Use

- **Organizational Data:** If you're part of an organization, its owners might see your activity.
 - **Aggregated Data:** We might use or share combined data that doesn't identify you.
 - **Contractor and Vendor Info:** We manage your work-related info for service agreements and legal purposes.
 - **Billing and Contracts:** We use data for billing, contracts, and payments.
 - **Public Data:** We can freely share data that doesn't identify individuals.
-

*Potential Risks and Sensitive Information

- **Sensitive Personal Data:** We might collect sensitive info like race, politics, religion, health, or sexuality.
 - **User-Generated Content:** Be cautious; your comments or uploads could be seen by others. We aren't responsible for their privacy.
-

Data Protection and Retention

- **Security Measures:** We protect your data from unauthorized access and misuse.
 - **Data Retention:** We only keep your data as long as needed and allowed by law.
-

Content You Create or Access

- **User Content:** We collect what you create or interact with, including photos, videos, comments, likes, and more.
-

Figure 8: Summarization by Kimi (Moonshot V1).

is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? (If so, please provide a description.)

No, all these privacy policies are publicly available.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** (If so, please describe why.)

No, the APPSI-139 does not include any data that could be considered offensive, insulting, threatening, or anxiety-provoking. This is because the dataset comprises privacy policies, and public documents to inform users and obtain their consent.

- **Does the dataset relate to people?** (If not, you may skip the remaining questions in this section.)

No, the APPSI-139 does not pertain to individuals or personal data. It focuses solely on the content and structure of privacy policies.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please

describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

N/A.

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**(If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.)

N/A.

- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** (If so, please provide a description.)

N/A.

- **Any other comments?**

Algorithm 2 Privacy Policy Summarization Instruction Template.

- 1: **Task Description:**
 - 2: You are an expert in privacy policies. Please summarize the following privacy policy text sentence by sentence and classify each according to the predefined categories of data practices.
 - 3: **Privacy Policy Categories:**
 - 4: *First-Party Collection:* The types of user information collected by the service provider, the purposes of collection, and whether it is mandatory for users to provide this information.
 - 5: *Permission Acquisition:* How the service provider obtains user permissions for the application, the purpose of these permissions, and whether these permissions are mandatory for users to grant.
 - 6: *Third-Party Sharing/Disclosure:* The purposes, types, and methods of sharing or disclosing user information to third parties, and the compliance with regulations.
 - 7: —*Other categories can be added as needed.*—
 - 8: **Special Markings:**
 - 9: *Risk:* Assigned to content that may raise compliance concerns, including potential violations of major data protection regulations (e.g., GDPR, CCPA) or vague descriptions of data handling practices.
 - 10: *Sensitivity:* Covers data practices involving users sensitive personal information, such as biometric data, precise geolocation, financial account details, etc.
 - 11: **Task Requirements:**
 - 12: 1. Read and analyze the privacy policy text sentence by sentence or paragraph by paragraph, identifying and categorizing each part according to the relevant data practices.
 - 13: 2. Summarize each data practice category using clear and simple language, ensuring that non-expert users can easily understand the core content.
 - 14: 3. Use a multi-level bullet point structure to organize the summary results, making it easier to review and analyze. Use a concise, indented list structure.
 - 15: 4. For clauses that involve Risk or Sensitivity, clearly mark them and briefly explain the potential issues or violations associated with them.
 - 16: **Please summarize the following privacy policy based on the above instructions:**
 - 17: [—*Insert privacy policy paragraph here*—]
-

N/A.

(How were these mechanisms or procedures validated?)

H.3 Collection Process

- **How was the data associated with each instance acquired?** (Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.)

Please check Section 3.1 in the main paper.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**

We have developed a straightforward web crawler application to systematically retrieve public privacy policies from the Google Play Store and the Apple App Store. This program ensures a streamlined and efficient process for accessing these essential documents.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

We utilized web crawling to collect English privacy policies from the top 100 apps in the U.S. on both the Google Play Store and Apple App Store. After deduplication and content verification, we obtained 139 complete application privacy policies, which were then segmented into sentence-level samples using regular expressions.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

All annotators received appropriate compensation according to standard practice, significantly above the local average wage.

- **Over what timeframe was the data collected? (Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** If not, please describe the timeframe in which the data associated with the instances was created.)

We collect privacy policy updates until October 2023.

- **Were any ethical review processes conducted (e.g., by an institutional review board)?** (If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.)

Yes. We have conducted an internal ethical review process by the ethical team.

- **Does the dataset relate to people?** (If not, you may skip the remaining questions in this section.)

No, the APPSI-139 does not pertain to individuals or personal data.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?** Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A.

- **Were the individuals in question notified about the data collection?** (If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.)

N/A.

- **Did the individuals in question consent to the collection and use of their data?** (If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.)

N/A.

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** (If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).)

N/A.

- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** (If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.)

N/A.

- **Any other comments?**

N/A.

H.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** (If so, please provide a description. If not, you may skip the remainder of the questions in this section.)

We removed the HTML tags because they are meaningless for privacy policy research.

- **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** (If so, please provide a link or other access point to the “raw” data.)

Yes. We have archived the “raw” data on GitHub, ensuring its accessibility and preservation for future reference and analysis.

- **Is the software used to preprocess/clean/label the instances available?** (If so, please provide a link or other access point.)

The entire annotation process was conducted using the Doccano (Hiroki et al., 2018) tool, which is accessible via the following link: <https://github.com/doccano/doccano>. Visual demonstrations of the annotation process are depicted in Figures 9 and 10. To foster transparency and reproducibility, we provide the source code, annotation guidelines, and dataset in a public repository, which can be accessed via <https://github.com/EnlightenedAI/APPSI-139>.

- **Any other comments?**

None.

H.5 Uses

- **Has the dataset been used for any tasks already?** (If so, please provide a description.)

The APPSI-139 dataset represents a novel resource for English application privacy policy summarization and interpretation. We have conducted extensive experiments to benchmark the performance of both classical machine learning algorithms and cutting-edge deep learning models across tasks such as automatic text summarization, text generation, and text classification. The comprehensive findings of these experiments are elaborated in Section 4 of our primary research document.

- **Is there a repository that links to any or all papers or systems that use the dataset?** (If so, please provide a link or other access point.)

We provide the source code, annotation guidelines, and dataset in a public repository, which can be accessed via <https://github.com/EnlightenedAI/APPSI-139>.

- **What (other) tasks could the dataset be used for?**

This dataset supports research and implementation across a variety of methods, including natural language classification and retrieval, text generation, in-context learning (including zero-shot and few-shot), and instruction

fine-tuning. Beyond labeling and retrieving privacy policy content, the dataset can also optimize large language models through in-context learning and instruction fine-tuning, adjusting summary content based on user needs and thus improving the model’s performance in privacy policy applications. The source code and dataset are provided in a public repository, which can be accessed via <https://github.com/EnlightenedAI/APPSI-139>.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** (For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?)

No.

- **Are there tasks for which the dataset should not be used?** (If so, please provide a description.)

No.

- **Any other comments?**

N/A.

H.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** (If so, please provide a description.)

No.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** (Does the dataset have a digital object identifier (DOI)?)

It is released on Github at <https://github.com/EnlightenedAI/APPSI-139>. No DOI.

- **When will the dataset be distributed?**

Before the conference.

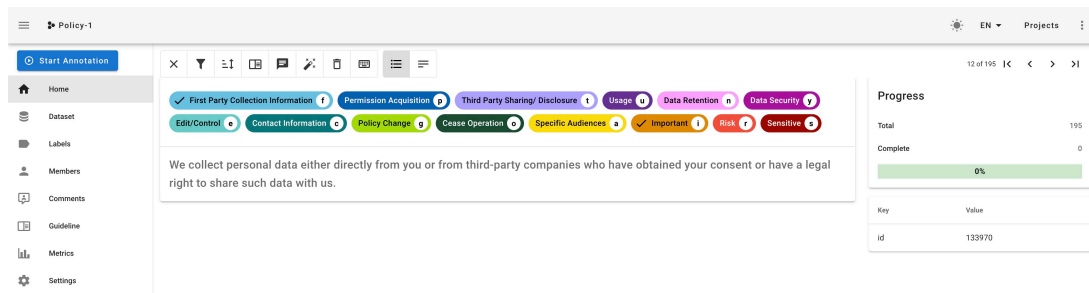


Figure 9: Annotation of Privacy Policy in Doccano

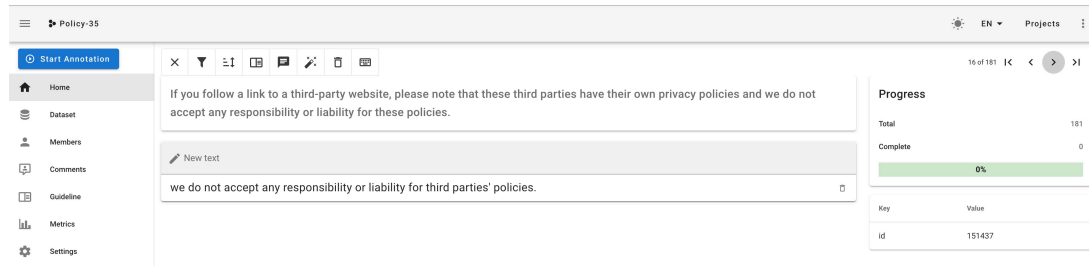


Figure 10: Rewritten of Privacy Policy in Doccano

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** (If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.)

CC BY 4.0. (<https://creativecommons.org/licenses/by/4.0/>)

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.)

No.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.)

No.

- **Any other comments?**

N/A.

H.7 Maintenance

- **Who is supporting/hosting/maintaining the dataset?**

This dataset was created by an interdisciplinary research group. *The specific details of the team and the institution are currently withheld due to anonymity policies.*

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

E-mail addresses are at the top of this document.

- **Is there an erratum?** (If so, please provide a link or other access point.)

No.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** (If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?)

No. If we plan to update the dataset in the future, we will elaborate on the reason on our GitHub repository.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g.,**

were individuals in question told that their data would be retained for a fixed period of time and then deleted)? (If so, please describe these limits and explain how they will be enforced.)

No.

- **Will older versions of the dataset continue to be supported/hosted/maintained?** (If so, please describe how. If not, please describe how its obsolescence will be communicated to users.)

Yes. If we plan to update the data, we will maintain the old version and then release the follow-up version, for example, APPSI-139-V2.0.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** (If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.)

Yes. For data annotation, researchers could carefully check our annotation guidelines in GitHub to ensure consistency. And if others want to contribute to the dataset, they could submit a pull request or contact us via email.

- **Any other comments?**

N/A.