TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214
DOI: 10.26599/TST.2022.90100
Volume xx, Number x, xxxxxxx 20xx

A Privacy Policy Text Compliance Reasoning Framework with Large Language Models for Healthcare Services

Jintao Chen, Fan Wang, Shengye Pang, Mingshuai Chen*, Meng Xi, Tiancheng Zhao, Jianwei Yin*

Abstract: The advancement of AI-generated content (AIGC) drives the diversification of healthcare services, resulting in increased private information collection by healthcare service providers. Therefore, compliance with privacy regulations has increasingly become a paramount concern for both regulatory authorities and consumers. Privacy policies are crucial for consumers to understand how their personal information is collected, stored, and processed. In this work, we propose a privacy policy text compliance reasoning framework called FACTOR, which harnesses the power of large language models (LLMs). Since the General Data Protection Regulation (GDPR) has broad applicability, this work selects GDPR Article 13 as regulation requirements. FACTOR segments the privacy policy text using a sliding window strategy and employs LLM-based text entailment to assess compliance for each segment. The framework then applies a rule-based ensemble approach to aggregate the entailment results for all regulation requirements from GDPR. Our experiments on a synthetic corpus of 388 privacy policies demonstrate the effectiveness of FACTOR. Additionally, we analyze 100 randomly selected websites offering healthcare services, revealing that 9 of them lack a privacy policy altogether, while 29 have privacy policy texts that fail to meet the regulation requirements.

Key words: Service Regulation, Privacy Policy, Compliance Reasoning, Healthcare Services

1 Introduction

The advancement of AI-Generated Content (AIGC) holds immense potential for smart healthcare services [1], such as automatic diagnosis [2] and healthcare prediction. The ability of AIGC allows these services to analyze vast amounts of data, including patient information and medical records. While enjoying the convenience brought by smart healthcare services, consumers are required to provide more and more private information which serves as paramount sources of innovation [3][4][5]. This trend renders certain healthcare service providers susceptible to security lapses in the acquisition, management, and interpretation of consumers' confidential information. Privacy policy is designed to inform consumers how to collect, process, store, and disclose their personal information. However, defining a compliant privacy policy is a difficult task [6]. Smart devices for healthcare frequently collect consumers' data to enable algorithmic decision-making processes [7], sometimes without providing fully informed consent [8]. As stated in Fig.1, users must provide the Studios with a permanent, irrevocable, and non-exclusive right to utilize the content and contributions they submit

• Jintao Chen, Fan Wang, Shengye Pang, Mingshuai Chen and Jianwei Yin are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China. E-mail: {chenjintao, fanwang97, pangsy, m.chen }@zju.edu.cn; zjuyjw@cs.zju.edu.cn.

• Meng Xi is with the School of Software Technology, Zhejiang University, Ningbo 315048, China. E-mail: mengxi@zju.edu.cn.

• Tiancheng Zhao is with the Binjiang Institute of Zhejiang University, Hangzhou, China. E-mail: tianchez@zju-bj.com

∗ To whom correspondence should be addressed.

Manuscript received: 2024-03-05; revised: 2024-05-01; accepted: 2024-05-08

Users take full responsibility for the content and entries they post. Users agree to release Studios from any legitimate claims from third parties arising from a culpable violation of the users' obligations. Studios explicitly does not claim content entered by users as its own. However, users shall grant Studios the permanent, irrevocable, non-exclusive right to use the content and contributions posted by the users.

Fig. 1 A sample privacy policy text exhibiting compliance infringements

is in contravention of the General Data Protection Regulation (GDPR). Specifically, this requirement infringes upon users' rights to rectify, erase, and object to the processing of their personal data as outlined in GDPR Article 13(2). These rights are fundamental to the GDPR, ensuring individuals have control over their personal information and the ability to request modifications, deletions, or restrictions on its usage. The policy's clause, by denying users the opportunity to withdraw or alter their data post-submission, disregards these essential rights. A comprehensive analysis of 79 health and wellness apps that have been certified as clinically safe and trustworthy by the United Kingdom National Health Service Health Apps Library revealed some concerning findings. The study indicated that 20% of the apps lacked any form of privacy policy, raising questions about the protection of consumers' data [9]. In addition, 78% of the policies did not specify the personal information contained in transmission [9].

Privacy policies are lengthy and complicated documents, making them difficult to read, infrequently reviewed, and not conducive to informed decision-making [10][11][12]. In practice, reading a complex privacy policy document takes prohibitive amounts of time, so it's rarely done by consumers [13][14]. This makes some consumers passively agree to healthcare services collecting privacy information in violation of regulations and laws. Consumers commonly tend to swiftly click the Agree button without thoroughly perusing the content of privacy policies [15].

Due to the emergence of violations such as privacy leaks, ensuring compliance with data privacy regulations is becoming an increasingly important societal issue. In the past decade, the implementation of the General Data Protection Regulation (GDPR) has catalyzed the most extensive modifications to privacy policies seen thus far [16][17]. The GDPR has 11 chapters, a total of 99 articles, of which article 13 in Chapter 3 clarifies information to be provided where personal data are collected from the users. Following the implementation of GDPR, a significant percentage of websites,

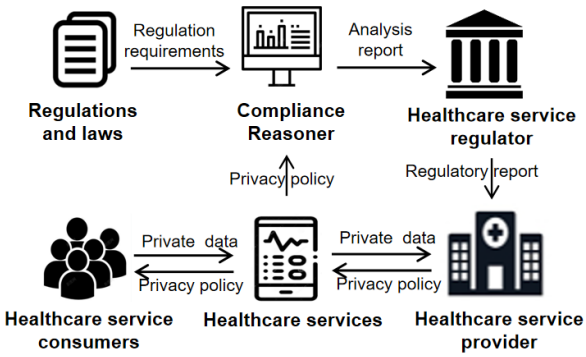

Fig. 2 Privacy policy regulation for smart healthcare services

approximately 72.6%, have proactively undertaken the task of revising and enhancing their privacy policies [18]. More specifically, websites that fall under the jurisdiction of the European Union (EU) have experienced a notable 35.39% increase in the textual length of their privacy policies [15]. Due to the wide applicability and impact of GDPR, this work targets privacy policy compliance analysis under the requirements of GDPR. The research findings [19][20] indicate that, despite advancements in GDPR enforcement, a considerable number of privacy policies examined do not fully meet the requirements set forth by the GDPR.

For healthcare service regulators and providers, systematically checking whether the privacy policy text is compliant requires a lot of effort, and it is unaffordable to rely on manual work. Consequently, automatic compliance analysis of privacy policies has attracted a lot of attention. As depicted in Fig. 2, healthcare providers publish applications on smart devices with privacy policy uploading. For example, mobile video streaming services for healthcare have attracted more and more consumers [21]. After healthcare service consumers agree to privacy policy, their private data is collected, processed, and stored by healthcare service providers. Due to the constant emergence and rapid iteration of healthcare services, the efficiency of the compliance analyzer is critical. Efficient privacy policy text compliance analytics are of utmost importance for healthcare service regulators to swiftly respond to privacy breaches. For privacy policy compliance analysis, the potential of Natural Language Processing (NLP) technologies has begun to be explored [22]. At present, the related works mainly cover two aspects which include privacy policy corpora construction and privacy policy text compliance analysis. The privacy policy corpora construction covers multiple languages such as English,

Chinese, and Arabic, and includes sentence-level and fine-grained phrase annotations. In addition to the privacy policy text corpora, there is also a corpus of 1,043 privacy laws, regulations, and guidelines covering 182 jurisdictions around the world [23]. For privacy policy text compliance analysis, there are research works that mainly include methods based on sentence classification [24] and methods based on text summarization [25]. However, the methods based on sentence classification make it difficult to exploit the contextual semantics of privacy policy text. Moreover, it also requires a lot of manpower for manual data annotation. The methods based on text summarization may lose some information in the complex privacy policy text which is critical for compliance analysis. As an excellent AIGC tool, large language models (LLMs) have shown extraordinary ability in text comprehension. In this work, we propose the FACTOR, a framework of privacy policy text compliance reasoning with LLMs. Despite the impressive performance in a variety of language tasks, LLMs are usually limited to processing text within the size of the context window. This work segments the privacy policy text into several paragraphs as input based on the sliding window strategy. This framework employs the text entailment task and a rule-based ensemble module to derive the reasoning results. Specifically, regulation requirements extracted from regulations and laws are treated as hypotheses. The paragraphs of a privacy policy are premises. Then the results of text entailment with all paragraphs will be ensembled based on defined rules to determine whether the privacy policy is compliant. The major contributions of this work are summarized as follows:

- A privacy policy text corpus with premise-hypothesis pairs for privacy policy compliance analysis is constructed based on the existing annotated corpora. The corpus takes GDPR Article 13 as the regulatory requirements which are summarized into 10 types of hypotheses.
- A framework for privacy policy text compliance reasoning named FACTOR is proposed which leverages contextual semantic information as much as possible. Taking premise-hypothesis pairs as input, the FACTOR implements privacy policy compliance reasoning based on text entailment.
- We experimentally evaluate the performance of FACTOR and the other two baselines to measure the effectiveness of our framework.
- To the best of our knowledge, the first publicly available corpus of privacy policies in healthcare services since the implementation of the GDPR is constructed. In addition, we conduct an analysis on this corpus which reveals that further refinement of certain aspects of some healthcare service privacy policies is necessary to align with regulation requirements.

The remainder of the paper is organized as fos:llows. Section 2 reviews the previous research about privacy policy corpora and privacy policy compliance analysis. We introduce the synthetic corpus for privacy policy compliance analysis in Section 3. In Section 4, we propose the FACTOR, a framework of privacy policy text compliance reasoning. Then we conduct experiments and analysis in Section 5. Finally, we conclude this work in Section 6.

2 Related Work

The development of service-oriented software engineering [26] promotes consumer concerns about private data. As more and more data compliance regulations and laws are introduced and enforced, tremendous efforts have been devoted to ensuring privacy compliance. For instance, federated learning has been widely used to fulfill privacy-aware requirements [27][28]. Research works have emerged to protect the privacy of data transmission and storage processes based on blockchain technology[29][30][31]. In this section, we review the literature related to privacy policy corpora and privacy policy compliance analysis.

2.1 Privacy Policy Corpora Creation.

Privacy policies serve as contractual agreements between service providers and users, outlining the terms and conditions regarding the collection, utilization, and disclosure of users' personal information by the companies. The earliest dataset, provided by Ramanath et al. in 2014, consisted of more than 1,000 privacy policies that were manually segmented [32]. Liu et al. [24] construct a corpus with sentence-level annotations based on 304 privacy policies. In terms of fine-grained annotations, the CA4P-483, a fine-grained Chinese privacy policy dataset is created in [33]. The dataset to facilitate fine-grained information extraction, namely PolicyIE, is presented in [34]. This English corpus consists of 5,250 intenr and 11,788 slot annotations over

Table 1 The comparison of commonly used privacy policy corpora

	No.Documents	Annotation granularity	Annotations	No.Labels	Language
Liu et al.	304	sentence-level	36610 sentences	11	English
PrivAud-100	100	sentence-level	3529 sentences	21	English
CA4P-483	483	fine-grained	11565 sentences	7	Chinese
OPP-115	115	fine-grained	102576 text spans	10	English
Ours	388	document-level	1263 violation issues	10	English

31 privacy policies. The APP-350 [35] is a corpus of human-annotated Android apps' privacy policies. For website privacy policy, Wilson et al. [36] create the corpus named OPP-115 which consists of 23K data practices, 128K practices attributes and 103K annotated text span. Furthermore, a dataset of connections between the OPP-115 annotation scheme and the principles and articles of the GDPR is created in [37]. Al-Khalifa et al. [38] introduce a Saudi Privacy Policy Dataset which is annotated according to the Personal Data Protection Law. Srinath et al. [39] introduced PrivaSeer, a dataset containing over 1 million English privacy policies extracted from the May 2019 Common Crawl dataset. For automatic detection of vague content in privacy policies, Lebanoff et al. [40] construct a sizable text corpus including human annotations for 133K words and 4.5K sentences. A comprehensive government privacy instruction corpus, comprising 1,043 privacy laws, regulations, and guidelines, has been introduced in [23]. The first bilingual corpus of mobile app privacy policies is introduced in [41]. This corpus contains 64 privacy policies in English and 91 privacy policies in German. The PrivacyQA, a challenging corpus constructed in [42], includes 1,750 questions related to the privacy policies of mobile applications and over 3,500 expert annotations of corresponding answers. In general, there are many annotated privacy policy corpora available by annotating and transforming the privacy policies. Due to inconsistent annotation standards and transformation purposes, these corpora are difficult to use in a uniform manner.

2.2 Privacy Policy Compliance Analysis.

To the best of our knowledge, privacy policy compliance analysis mainly includes blockchain-based analysis, AI-based analysis, knowledge representation-based analysis, and code detection-based analysis.

Blockchain-based Analysis. Barati et al. [43] transfer GDPR rules to opcodes in smart contracts to verify the operations of providers on user data. Truong et al. [44] develop a GDPR-compliant personal data management platform which implemented on top of the Hyperledger Fabric permissioned blockchain framework. A framework based on blockchain and the Internet of Vehicles oriented to securing GDPR compliance is proposed in [45]. As for verifying GDPR compliance in the multi-cloud environment, [46] introduces a blockchain-centric and user-centric framework for data management in a cloud environment to facilitate GDPR-compliant data operations through well-defined smart contracts.

AI-based analysis. Torre et al. [47] devise an AI-assisted method for automatically classifying the content of a given privacy policy to check whether it meets the information requirements stipulated by GDPR. The sentence classification method based on the large model is used for compliance analysis in [24]. Ravichander et al. [22] highlight the importance of NLP for privacy policy compliance analysis. Lebanoff et al. [40] first adopted the generative adversarial network to detect vague content in privacy policies. An automated analysis framework is proposed in [25] which leverages the ability of BiLSTM multi-class classification and a BERT extractive summarizer.

Knowledge Representation-based Analysis. Knowledge representation methods are also applied to privacy policy compliance analysis. A. Bonatti et al. [3] encode fragments of the GDPR into a fragment of OWL2 in order to reduce the compliance checking and policy validation to subsumption checking and concept consistency checking. The OPPO, an upper-level ontological model, is specifically developed to effectively encode the data practices documented in the privacy policies of online social networks (OSNs)[48]. Tauqeer et al. [20] develop a knowledge graph-based tool for GDPR contract compliance verification. As a type of knowledge graph designed to capture statements within a privacy policy by representing them as relationships between various sections of the text, the PoliGraph is introduced in [49].

Code Detection-based Analysis. To further analyze whether the application complies with the GDPR in the actual execution process, research on privacy policy compliance based on code analysis is carried out [50][51].

3 The Synthetic Corpus for Privacy Policy Text Compliance Analysis

Existing works have invested a lot of effort in data annotation, providing high-quality data for privacy policy compliance analysis. However, due to the inconsistent compliance basis selected, the annotation granularity and results are also inconsistent. We select commonly used corpora with GDPR as the basis for compliance annotation, summarised in Table 1. These corpora cover privacy policy texts for both mobile and web services. As illustrated in Table 1, the annotation granularity of corpora contains sentence-level and fine-grained. Different corpora have varying definitions and scopes for their labels. Some corpora such as [33] cover multiple articles of the GDPR, while others [52][53] specifically annotate based on Article 13 of the GDPR.

Given the high cost of data annotation, utilizing existing annotation corpora to their fullest potential is a critical consideration. Therefore we select the corpora with labels covering GDPR Article 13 proposed by Liu et al. and PrivAud-100 to synthesize a corpus with a larger scale. Liu et al. [24]constructed a corpus that encompasses 11 distinct label categories, which include Collect Personal Information, Data Retention Period, Data Processing Purposes, Contact Details, Right to Access, Right to Rectify or Erase, Right to Restrict Processing, Right to Object to Processing, Right to Lodge a Complaint, Right to Data Portability and Other. The privacy policies within the corpus span across 22 application categories, such as Communication, Game, and Business. For the annotation process, a group comprising legal and computer science experts has meticulously annotated these privacy policies, systematically identifying and categorizing diverse privacy-related aspects and concerns within each policy.

The PrivAud-100 constructed by [52], which consists of 100 randomly selected privacy policies, has 21 label categories. Eleven of the labels are the same as those mentioned above, leaving 10 labels containing Collect Health information, Collect Financial and payment information, Collect Authentication information, Collect Personal communications, Collect Location, Collect Web history, Collect User activity, Collect Website content, Collect Cookie and Secure Data Transfer. While the corpus may not have been annotated by experts, the authors utilized the Percent Agreement to assess the inter-rater reliability between them. The results indicate a strong agreement (0.96) across all annotated sentences. In this section we do not change the annotation of the original corpora. Thus, the quality of the annotation is maintained as in the original work.

The effective utilization of high-quality annotations within both corpora facilitates the analysis of privacy policy compliance. In this work, we synthesize a larger corpus based on these two corpora to enable fine-tuning of LLMs. The synthesis process consists of the following steps:

Unify the corpora labels. The corpus proposed by Liu et al. contains 11 labels extracted from Article 13 of GDPR. Among them, the 10 labels except *Other* correspond to the specific terms in Article 13 as illustrated in Table 2. For the PrivAud-100, its 11 tags are the same as the corpus proposed by Liu et al., and the other 10 labels refer to collecting specific user information, which can be converted into *Collect Personal Information*. Since *Other* cannot correspond to the requirements of Article 13 and has no role in compliance analysis, we did not consider it in the last step to generate hypotheses.

Rule-based violation annotation. The clauses of Article 13 follow the pattern that if the service provider collects the information of users, the required information must be provided by the service provider. Therefore, we perform document-level automatic annotation according to the 9 rules proposed in [24]. Since both corpora are annotated at the sentence level, we stitch sentences with consistent provenance into a single text. This text comes with a set containing the labels of all the sentences. We then apply the defined 9 rules for automated compliance labeling of the text. If a privacy policy text complies with these 9 rules, it will be judged automatically as compliant. Those that don't meet these 9 rules are automatically flagged as violations, and the reasons for the violation will be marked. In this step, we find that a violated privacy policy text is often the result of multiple violations. So far, we have constructed a corpus that supports privacy policy text compliance determination and traceability analysis of the causes of violations. Since the original corpora guarantees high-quality labeling, the synthetic corpus with logical operation inherits this property.

After the above synthesis steps, the synthetic corpus

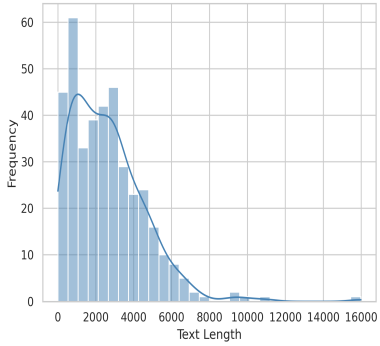

Fig. 3 The text length distribution of privacy policies

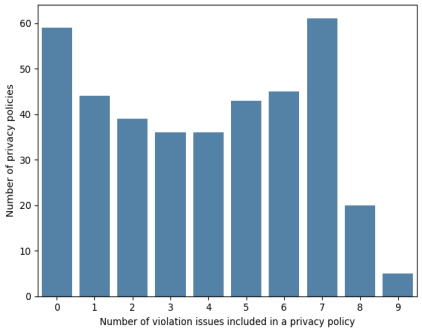

Fig. 4 Distribution of violation issues of privacy policies

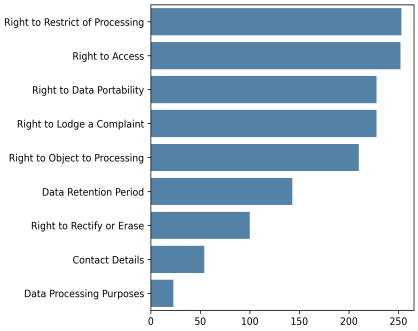

Fig. 5 The frequency of different violation issues

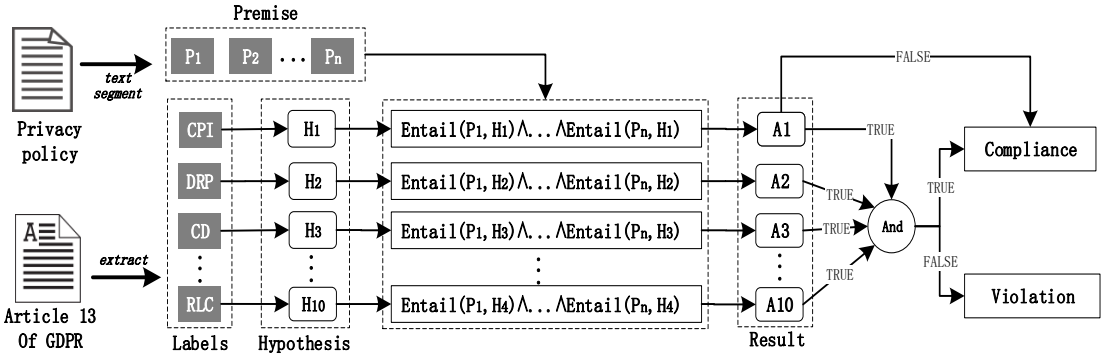

Fig. 6 The framework of privacy policy text compliance reasoning

has 388 privacy policy documents. The average text length of a privacy policy is 2596 words. The text length distribution of privacy policies in the corpus is shown in Fig. 3. Among them, 308 privacy policies have 1263 violation issues. The remaining 59 privacy policies comply with the requirements of Article 13 and are marked as compliance. Fig. 4 reveals the fact that the non-compliant privacy policy may contain multiple violations. There are 285 privacy policies with greater than or equal to 2 violation issues, representing 73.45% of the corpus. The number of privacy policies containing 7 violation issues is the highest at 65, representing 15.72% of the corpus. Therefore, it is challenging to correctly identify all the violation issues. In Fig. 5, it can be found that the violations of *Right to Restrict of Processing* and *Right to Access* are the two most frequent violation types. Non-compliance incidents that violate the requirements for *Data Processing Purposes* occurred the least frequently.

4 The Framework for Privacy Policy Text Compliance Reasoning

Existing works for privacy policy text compliance reasoning mainly include methods based on sentence classification and methods based on text summarization. The methods based on sentence classification have difficulty in exploiting contextual semantics. As a result, some critical information will be lost. For methods based on text summarization, there is a lack of high-quality and sizeable datasets of textual summaries of privacy policies for compliance analysis. Therefore, we propose a framework named FACTOR which exploits contextual semantics while preserving all textual information through textual entailment.

4.1 The Definition of Privacy Policy Text Compliance Reasoning Task

The privacy policy text compliance reasoning task in this work is to judge whether the given privacy policy text is compliant based on the clauses of Article 13 of GDPR. Both the privacy policy and Article 13 are in natural language text. The long text of the privacy pol-

icy and professional requirements for the interpretation of GDPR clauses make compliance reasoning very challenging. The 10 labels in Table 2 summarize the compliance requirements of Article 13 for privacy policies that collect personal information. This reduces the difficulty of interpretation of GDPR clauses for language models. Based on these 10 labels, we constructed 10 hypotheses, as long as the privacy policy can meet these 10 hypotheses, then it is compliant. That is, for each privacy policy text T, it has a set of hypotheses H:

$$H = \{h_1, h_2, ..., h_i\}, \quad 1 \leq i \leq 10 \tag{1}$$

The premise set P is obtained by segmenting the privacy policy text T:

$$P = \{p_1, p_2, ..., p_x\}, \quad 1 \leq x \leq n; \tag{2}$$

where n is the number of paragraphs obtained by the segmentation of T. Due to the inconsistent length of the privacy policy text, the value of n is also not fixed. For the hypothesis based on the *Collect Personal Information* label, given the premise obtained from the segmentation of a certain privacy policy text, if there are one or more premises that can be inferred to entail the hypothesis, then the privacy policy text meets the regulatory requirements expressed by the label. Then, the privacy policy text compliance reasoning task can be defined as: for a privacy policy text T, given its premises set P and the hypotheses set H,

$$l_i = TE(p_1, h_i) \vee TE(p_2, h_i) \vee ... \vee TE(p_x, h_i), \\ 1 \leq x \leq n, \quad 1 \leq i \leq 10 \tag{3}$$

where the l_i is the label of whether the hypothesis h_i can be inferred from the privacy policy text. The $TE(p_x, h_i)$ refers to the textual entailment function which takes a premise-hypothesis pair as input. The output of $TE(p_x, h_i)$ is whether the hypothesis h_i is entailed by the premise p_x. As long as there is a promise p_x that can be deduced to get h_i, then the label l_i is TRUE. Based on the label set L, we use the following formula to judge whether the privacy policy is compliant:

$$R = \neg l_1 \vee (l_1 \wedge l_2 \wedge ... \wedge l_{10}) \tag{4}$$

where R is the compliance label of the privacy policy text. If the value is TRUE, it means that the privacy policy is compliant, and if it is FALSE, it is in violation.

4.2 Overview of the FACTOR

Fig. 5 is an overview of the FACTOR. It takes as input the privacy policy text and GDPR Article 13. As the privacy policy text can be very lengthy and complex, the framework segments the privacy policy text into several paragraphs as premises, each containing no more than a certain number of words. The regulation requirements of GDPR Article 13 are summarized into corresponding hypotheses through 10 labels (acronym in Fig. 6). The hypotheses are listed in Table 2. The premise-hypothesis pairs are then fed into the textual entailment reasoner. Based on the output of the reasoner, the final reasoning result is obtained according to formula 4. The privacy policy compliance text reasoning is described in Algorithm 1.

Algorithm 1 Compliance reasoning

Input: a privacy policy text T; hypotheses set H extracted from GDPR Article 13;
the length of premise α; the step size β.
Output: result R of compliance reasoning.

$P \leftarrow \emptyset, L \leftarrow \emptyset$
$P \leftarrow getPremise(T, \alpha, \beta)$ {//Segment text into multiple premises }
for $h_i \in H$ **do**
  $l_i \leftarrow False$
  for $each \quad p \in P$ **do**
    if $TE(p, h_i) == True$ **then**
      $l_i \leftarrow True$ {//Conduct textual entailment for premise-hypothesis pairs }
    end if
  end for
  $L.add(l_i)$ {//L is a label set whether the hypotheses entailed by any premise}
end for
for $l_i \in L$ **do**
  $R \leftarrow True$
  if i>1 $\quad and \quad l_i == False$ **then**
    $R \leftarrow False$ {//Determine whether the privacy policy text is compliant}
  end if
end for
return R

4.3 Sentence Segmentation with Sliding Windows for Premises.

The length of the privacy policy text is too long to directly serve as the premise, and the complicated semantics of the entire privacy policy text will increase the difficulty of text entailment. Therefore, we divide the entire privacy policy text based on the sliding window strategy to obtain multiple promises. In order to ensure that the premise obtained after segmentation has complete sentence semantics, the unit we process is not a word but a sentence. The function $getPresmise()$ is defined in Algorithm 2. It first converts the entire text

Algorithm 2 Text segmentation for premises

Input: a privacy policy text T; the length of premise α; the step size β.

Output: A set of premises P obtained by segmenting the privacy policy text.

$def \quad getPremise(T, \alpha, \beta)$:
  $P \leftarrow \emptyset$
  $S \leftarrow List(T.sentences)$ {//Split the privacy policy text into sentences }
  $n \leftarrow len(S), start \leftarrow 0$
  while $start < n$ **do**
    $end \leftarrow start + \beta$ {//Segmentation in units of sentences}
    if $end > n$ **then**
      $end \leftarrow n$ {//Make sure the sentence list length is not exceeded}
    end if
    $text \leftarrow S[start : end]$
    while $len(text) < \alpha$ **do** {//Premise length is less than α}
      $end \leftarrow end + 1$
      $text \leftarrow S[start : end]$
    end while
    $P.add(text)$
    $start \leftarrow start + \beta$
  end while
  return P

into a list of sentences S. Based on the given length α of a premise, the function starts at the beginning of S and slides the window at a fixed step size β processing the sentences in the window as a new text segment each time until it reaches the end of the S. In this way, a long privacy policy text can be divided into several overlapping text fragments as premises.

4.4 Text Entailment for Privacy Policy Text Compliance Reasoning.

Text entailment is one of the tasks of natural language inference. Given a premise text and a hypothesis text, it infers the relationship between the text pair according to the premise. There are three types of inference results: entailed, contradictory, and neutral. In the reasoning of privacy policy text compliance, GDPR Article 13 stipulates that service providers must provide users with corresponding information when collecting personal information. Therefore, only the entailment case is legal. The contradiction and neutral cases are violations. That is, in the case of collecting personal information, the privacy policy text is only compliant if every hypothesis is entailed by the privacy policy text. Since a privacy policy text is segmented into multiple premises, for each hypothesis h_i, there exists any promise $p_x \in P$ that makes h_i is entailed in p_x, then h_i is entailed by the privacy policy text T. Since the LLMs are trained on a large amount of text data and have achieved impressive performance on various NLP tasks, this work selects the LLMs as the reasoner.

5 The Experiments and Discussion

To measure the effectiveness of our framework, we conduct experiments and ablation studies on the synthetic corpus. Simultaneously, we randomly crawl the privacy policies of 100 healthcare services to construct a public corpus. To the best of our knowledge, this is the first publicly available corpus of healthcare services privacy policies since the implementation of the GDPR. We apply the FACTOR to this corpus and find that the issue of healthcare services privacy policy compliance requires more effort.

5.1 Experiment Setup

The RoBERTa [54], a pre-training model, has achieved excellent performance on multiple natural language processing tasks. Therefore, we select the RoBERTa-base of the Huggingface * to achieve privacy policy compliance reasoning. The RoBERTa-base model has a total of approximately 125 million parameters. The model size of RoBERTa-base consists of 12 layers of the Transformer architecture, with each layer having a hidden size of 768 units. Our experimental setup is designed to operate under a Zero-shot condition. Since the maximum text length that RoBERTa-base can handle is 512, and the hypothetical text length is less than 30 words, we conduct experiments when the premise lengths were set to 350, 400, and 450, respectively. Moreover, we test the impact of different sliding window step sizes on the performance. Finally, we measure the performance of directly segmenting the privacy policy text into multiple premises for ablation analysis to confirm that the sliding window method works.

Hypotheses generation. The privacy policy text compliance reasoning is to infer whether a given privacy policy text satisfies regulation requirements. Based on the 10 categories of labels extracted from Article 13, we summarize the corresponding hypotheses listed in Table 2. The clauses corresponding to labels are also indicated in the Table 2. If regulation requirements change, service regulators can quickly locate the appropriate label and make adjustments. For each privacy policy text,

*https://huggingface.co/roberta-base/tree/main

Table 2 The labels in the corpus with corresponding clauses and hypotheses

Labels	Clauses	Hypotheses
Collect Personal Information	*Art 13.1*	The privacy policy context specifies the collection of personal information.
Data Retention Period	*Art 13.2(a)*	The privacy policy context specifies the period for which the personal data will be stored.
Data Processing Purposes	*Art 13.1(c)*	The privacy policy context specifies the purposes of the processing for which the personal data are intended.
Contact Details	*Art 13.1(a)(b)*	Does the privacy policy context specify the contact details of the data controller or the data protection officer.
Right to Access	*Art 13.2(b)*	The privacy policy context specifies the right of data subjects to request from the data controller to access their personal information.
Right to Rectify or Erase	*Art 13.2(b)*	The privacy policy context specifies the right of data subjects to request from the data controller to rectify or erase of their personal information.
Right to Restrict of Processing	*Art 13.2(b)*	The privacy policy context specifies the right of data subjects to request from the data controller to restrict processing concerning the data subjects.
Right to Object to Processing	*Art 13.2(b)*	The privacy policy context specifies the right of data subjects to request from the data controller to object to processing.
Right to Data Portability	*Art 13.2(b)*	The privacy policy context specifies the right of data subjects to receive and transmit his/her personal data to another data controller.
Right to Lodge a Complaint	*Art 13.2(d)*	The privacy policy context specifies the right of data subject to lodge a complaint with a supervisory authority.

we add these 10 hypotheses and mark whether each hypothesis is entailed by the privacy policy text according to the existing annotations.

Baseline selection. We choose two algorithms as our baselines to verify the effectiveness of the proposed framework:

- Sentence classification-based method. For sentence classification, the label *Other* still needs to be considered. Based on the given 11 labels, classify each sentence in the privacy policy text, and then perform compliance reasoning according to formula 4.
- Text summarization-based method. Summarize the privacy policy text, and perform multi-label classification on the sentences in the summary. The compliance reasoning is performed according to formula 4.

5.2 Experiment Results on Synthetic Corpus

As illustrated in Algorithm 1, we first segment the input text T and then implement privacy policy compliance reasoning based on textual entailment. To solve the problem caused by the long text of the privacy policy, we divide it into multiple input paragraphs as premises set P based on the sliding window. In order to ensure the semantics of the context, we use sentences as units in the segmentation process, so as to ensure that each premise is composed of complete sentences. The step size in the sliding window is also based on sentences as the unit of operation. We conduct experiments with premise lengths of 350, 400, and 450, and step sizes of 2 and 3, respectively. Table 3 is the experimental results of privacy policy compliance reasoning. We have calculated four metrics to evaluate performance: accuracy, F1 score, precision, and recall. Since the non-compliant privacy policy text accounts for 84.79%, and it is more costly to misinfer the non-compliant privacy policy text as a compliant privacy policy, we select the weighted F1-score for a detailed analysis.

When the step size is 2 and the premise length is less than 400, the accuracy and F1-score achieve the best performance. The possible reason is that when the step size is 3, each premise introduces more redundant information, which affects the reasoning results of text entailment. When the step size is fixed, the accuracy reaches the highest when the premise length is less than 400. A shorter premise length may weaken the semantics of the context, while a longer premise length is more likely to contain confusing text, making model reasoning more errors.

In order to further analyze the factors that affect the accuracy of privacy policy text compliance reasoning, we analyze the impact of different labels and text

Table 3 The results of privacy policy compliance text reasoning

		Len(Premise) <350	Len(Premise)<400	Len(Premise)<450
Step size = 2	No.Premises	3297	2903	2599
	No.RPP	264	286	256
	Accuracy	68.04%	73.71%	65.98%
	F1-score	72.67%	75.08%	69.47%
	Precision	75.15%	76.70%	74.23%
	Recall	70.62%	73.70%	65.98%
Step size = 3	No.Premises	3200	2844	2552
	No.RPP	262	276	254
	Accuracy	67.53%	71.13%	65.46%
	F1-score	70.59%	73.26%	69.01%
	Precision	74.63%	75.93%	73.78%
	Recall	67.53%	71.13%	65.46%

The No.RPP represents the number of privacy policies that are reasoned correctly. The NO.premises indicates the number of premises segmented under different settings. The Len(Premise) refers to the number of words in the premise.

lengths on the accuracy of compliance reasoning. Fig. 7 is the result of label reasoning accuracy except for the label *Other*. The *Collect Personal Information* label achieves an accuracy of 82%, while the reasoning accuracy for the *Data Processing Purpose* label is only 31%. We speculate that different service providers may have different data processing purposes, and the writing methods in the process of writing privacy policy texts are also ever-changing, which has caused difficulties for LLMs to understand. Compared with *Data processing purposes*, labels such as *Right to rectify or erase* and *Right to access* are better understood because of their clear content. Fig. 8 shows the results of compliance reasoning accuracy under different privacy policy text lengths. It can be seen that as the length of the text becomes longer, the accuracy of compliance reasoning decreases. Due to the small number of privacy policies with about 10,000 words, the accuracy fluctuates relatively large in the last few text length intervals, but the overall decline trend of the reasoning accuracy is still obvious.

Table 4 The comparison results

	Accuracy	F1-Score
FACTOR(with sliding window)	73.71%	75.08%
FACTOR(without sliding window)	62.63%	64.63%
Sentence classification-base	67.26%	67.56%
Text summarization-based	54.79%	59.26%

To demonstrate the effectiveness of jointly using both textual entailment and sliding window, we conduct comparative experiments based on sentence classification and text summarization respectively. Moreover, we compare the effect of using the sliding window strategy and not applying it. Table 4 is the results of comparison experiments and ablation studies. Table 4 shows that the application of the sliding window strategy can achieve an accuracy rate of 71.91%, which is 9.28% higher than that of the non-application sliding window strategy. Its F1-score is 9.18% higher, indicating that the application of the sliding window strategy makes the privacy policy text compliance reasoning framework more robust. According to the results in Table 4, in the privacy policy text compliance reasoning, the method based on text entailment has a higher accuracy than sentence classification and text summarization, which are 6.45% and 18.92% higher, respectively. Possible reasons for this result include:

- Sentence classification-based method segments the entire privacy policy text into sentences and then classifies the sentences. In the process of segmentation, the contextual semantics of the privacy policy text will be missing, resulting in lower accuracy of sentence classification.

- Text summarization-based method needs to divide long privacy policy text into paragraphs of no more than 512 words, and then summarize these paragraphs separately. All generated summary sentences are classified, and then compliance reasoning is performed based on formula 4. Because the summarization process is zero-shot, a lot of semantic information may be lost, resulting in low accuracy of inference results.

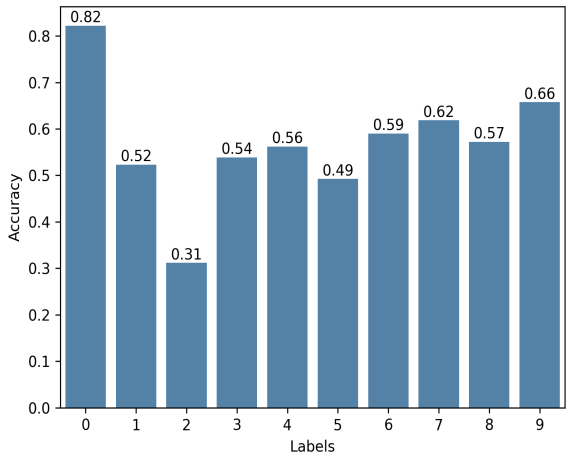

Fig. 7 Reasoning accuracies of labels

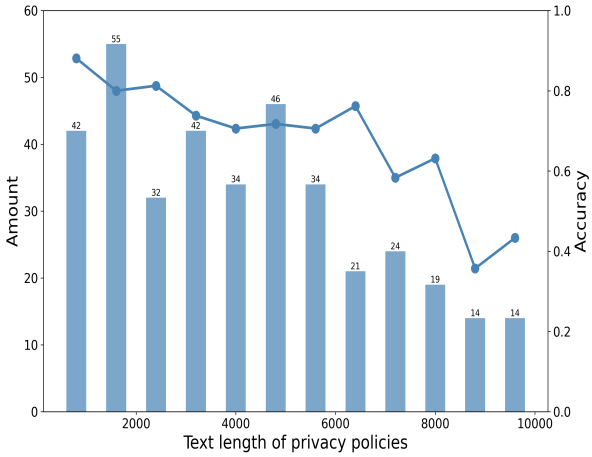

Fig. 8 Compliance reasoning accuracies under different text lengths

5.3 Analysis on Privacy Policies of Healthcare Services

Since the implementation of GDPR, many healthcare services have changed their privacy policies. The explosive growth of web services [55] makes automated analysis for privacy policy text urgent. However, to the best of our knowledge, there has been no publicly available corpus of healthcare service privacy policies used for compliance analysis since then. To facilitate privacy compliance, this work randomly crawls the privacy policies of 100 healthcare websites. Special characters and blank lines have been removed. Of these, nine websites have empty privacy policies. The average number of words in the privacy policy text for the remaining 91 applications is 2059. Out of these, 44 privacy policies contain child-oriented content. The distribution of privacy policy text length is depicted in Fig. 9. It can be observed that there are fewer distributions with child-oriented privacy policy text of less than 1,000 words. Moreover, the majority of the privacy policy text falls within the range of 1,000 to 2,000 words. The distribution of text length in healthcare services privacy policies bears resemblance to the synthetic corpus. We apply our framework to this corpus to analyze the compliance of healthcare service privacy policies. The 9 websites with empty privacy policies are directly judged to be in violation. Of the remaining 91 privacy policies, 29 are found to be in violation. As a result, the privacy policies of 100 randomly crawled healthcare services had a compliance rate of 62%. Moreover, only 44% of healthcare service privacy policies take children as a group into

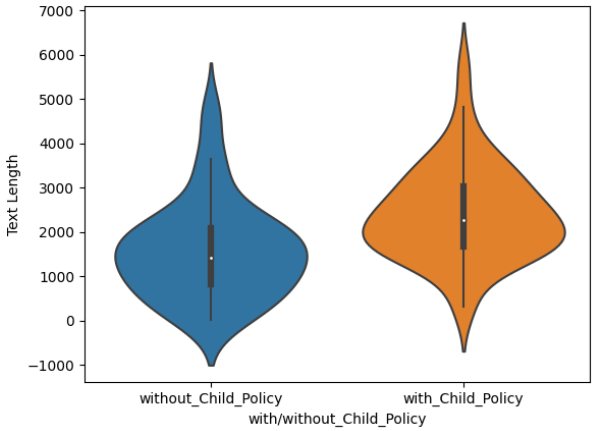

Fig. 9 The text length distribution of healthcare services privacy policies

account. And 13 of the privacy policies that take children into account are found to be in violation. In the non-empty violation privacy policy text, the violation issues centered on three main aspects: a) failure to state the period for which the personal data will be stored by healthcare service providers; b) failure to state the purpose of the processing for which the personal data are intended; and c) failure to state the right of data subject to lodge a complaint with a regulation authority.

Protecting children's information is crucial to ensure their safety, maintain their privacy, and shield them from potential harm, exploitation, and identity theft in the digital world. Therefore, we have conducted a further analysis based on the Children's Online Pri-

vacy Protection Act (COPPA) which is a federal law designed to protect the privacy of children under 13. As stipulated by the COPPA, children's information is federally protected, prohibiting the collection of any personal data from children through online platforms and digital connected devices. In compliance with the COPPA, we analyze 44 privacy policy texts that reference children and identify 34 privacy policies that explicitly state they do not collect information from children under the age of 13. The compliance rate among these 44 privacy policy texts is 77.27%. It underscores the necessity for continued oversight and enforcement to ensure that all privacy policies are in full compliance with COPPA, safeguarding the privacy rights of children under 13.

This analysis suggests that, in practice, there is still a need for improvement in healthcare service privacy policies to ensure compliance with regulation requirements.

6 Conclusions

The development of AIGC plays a crucial role in driving innovation in healthcare services. While there may be a temptation to access a vast amount of private information through devices and websites, it is of utmost importance to prioritize responsible data collection, processing, and storage, and to ensure the highest level of privacy protection. In this work, we propose a privacy policy text compliance reasoning framework with the help of LLMs. The framework achieves 73.71% compliance reasoning accuracy on the synthetic corpus. Furthermore, when applying the framework to the corpus of healthcare service privacy policies, it is revealed that 38% of healthcare service do not provide a privacy policy that is compliant with regulations. In future work, we are devoted to deeply investigating the following two aspects: (1) Improve the ability of LLMs to better understand regulation requirements for privacy policy of healthcare devices and websites. GDPR is a regulatory act with a large scope of influence. However, there are diverse laws and regulations in different countries and regions. How to further improve the understanding ability of LLMs based on deep correlation mining [56][57] between these laws and regulations is an important direction. (2) Conduct further analysis of the compliance of child-related content in the healthcare service privacy policy text. Considering their vulnerability and the sensitivity surrounding their personal information, preserving the privacy of children is of tremendous significance.

Acknowledgment

This work has been partially funded by the National Key R&D Program of China under grant No. 2022YFF0902600, by the ZJNSF Major Program under grant No. LD24F020013 and by the ZJU Education Foundation's Qizhen Talent program.

References

[1] X. Zhou, X. Zheng, T. Shu, W. Liang, I. Kevin, K. Wang, L. Qi, S. Shimizu, and Q. Jin, "Information theoretic learning-enhanced dual-generative adversarial networks with causal representation for robust ood generalization," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[2] P. Maji, H. K. Mondal, A. P. Roy, S. Poddar, and S. P. Mohanty, "ikardo: An intelligent ecg device for automatic critical beat identification for smart healthcare," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 4, pp. 235–243, 2021.

[3] P. A. Bonatti, L. Ioffredo, I. M. Petrova, L. Sauro, and I. R. Siahaan, "Real-time reasoning in owl2 for gdpr compliance," *Artificial Intelligence*, vol. 289, p. 103389, 2020.

[4] X. Zhou, W. Liang, K. Yan, W. Li, I. Kevin, K. Wang, J. Ma, and Q. Jin, "Edge-enabled two-stage scheduling based on deep reinforcement learning for internet of everything," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3295–3304, 2022.

[5] L. Kong, G. Li, W. Rafique, S. Shen, Q. He, M. R. Khosravi, R. Wang, and L. Qi, "Time-aware missing healthcare data prediction based on arima model," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.

[6] J. Leicht, M. Heisel, and A. Gerl, "Pripocog: Guiding policy authors to define gdpr-compliant privacy policies," in *International Conference on Trust and Privacy in Digital Business*. Springer, 2022, pp. 1–16.

[7] X. Zhou, X. Ye, I. Kevin, K. Wang, W. Liang, N. K. C. Nair, S. Shimizu, Z. Yan, and Q. Jin, "Hierarchical federated learning with social context clustering-based participant selection for internet of medical things applications," *IEEE Transactions on Computational Social Systems*, 2023.

[8] A. Bowyer, J. Holt, J. Go Jefferies, R. Wilson, D. Kirk, and J. David Smeddinck, "Human-gdpr interaction: practical experiences of accessing personal data," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–19.

[9] K. Huckvale, J. T. Prieto, M. Tilney, P.-J. Benghozi, and J. Car, "Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment," *BMC medicine*, vol. 13, no. 1, pp. 1–13, 2015.

[10] A. M. McDonald and L. F. Cranor, "The cost of reading privacy policies," *Isjlp*, vol. 4, p. 543, 2008.

[11] B. Fabian, T. Ermakova, and T. Lentz, "Large-scale readability analysis of privacy policies," in *Proceedings of the international conference on web intelligence*, 2017, pp. 18–25.

[12] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: Automated analysis and presentation of privacy policies using deep learning," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 531–548.

[13] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: a technological perspective and review. j big data. 2016."

[14] A. Gerl, "Modelling of a privacy language and efficient policy-based de-identification," Ph.D. dissertation, Université de Lyon; Universität Passau (Deutscheland), 2019.

[15] C. Tang, Z. Liu, C. Ma, Z. Wu, Y. Li, W. Liu, D. Zhu, Q. Li, X. Li, T. Liu *et al.*, "Policygpt: Automated analysis of privacy policies with large language models," *arXiv preprint arXiv:2309.10238*, 2023.

[16] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. Mayer, "Privacy policies over time: Curation and analysis of a million-document dataset," in *Proceedings of the Web Conference 2021*, 2021, pp. 2165–2176.

[17] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, "The privacy policy landscape after the gdpr," *Proceedings on Privacy Enhancing Technologies*, vol. 1, pp. 47–64, 2020.

[18] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz, "We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy," *Informatik Spektrum*, vol. 42, pp. 345–346, 2019.

[19] N. Bateni, J. Kaur, R. Dara, and F. Song, "Content analysis of privacy policies before and after gdpr," in *2022 19th Annual International Conference on Privacy, Security & Trust (PST)*. IEEE, 2022, pp. 1–9.

[20] A. Tauqeer, A. Kurteva, T. R. Chhetri, A. Ahmeti, and A. Fensel, "Automated gdpr contract compliance verification using knowledge graphs," *Information*, vol. 13, no. 10, p. 447, 2022.

[21] L. Qi, X. Xu, X. Wu, Q. Ni, Y. Yuan, and X. Zhang, "Digital-twin-enabled 6g mobile network video streaming using mobile crowdsourcing," *IEEE Journal on Selected Areas in Communications*, 2023.

[22] A. Ravichander, A. W. Black, T. Norton, S. Wilson, and N. Sadeh, "Breaking down walls of text: How can nlp benefit consumer privacy?" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 1, 2021.

[23] S. Gupta, E. Poplavska, N. O'Toole, S. Arora, T. Norton, N. Sadeh, and S. Wilson, "Creation and analysis of an international corpus of privacy laws," 2022.

[24] S. Liu, B. Zhao, R. Guo, G. Meng, F. Zhang, and M. Zhang, "Have you been properly notified? automatic compliance analysis of privacy policy text with gdpr article 13," in *Proceedings of the Web Conference 2021*, 2021, pp. 2154–2164.

[25] L. Elluri, S. S. L. Chukkapalli, K. P. Joshi, T. Finin, and A. Joshi, "A bert based approach to measure web services policies compliance with gdpr," *IEEE Access*, vol. 9, pp. 148 004–148 016, 2021.

[26] F. Wang, L. Wang, G. Li, Y. Wang, C. Lv, and L. Qi, "Edge-cloud-enabled matrix factorization for diversified apis recommendation in mashup creation," *World Wide Web*, pp. 1–21, 2021.

[27] X. Zhou, W. Liang, I. Kevin, K. Wang, Z. Yan, L. T. Yang, W. Wei, J. Ma, and Q. Jin, "Decentralized p2p federated learning for privacy-preserving and resilient mobile robotic systems," *IEEE Wireless Communications*, vol. 30, no. 2, pp. 82–89, 2023.

[28] X. Zhou, X. Zheng, X. Cui, J. Shi, W. Liang, Z. Yan, L. T. Yang, S. Shimizu, I. Kevin, and K. Wang, "Digital twin enhanced federated reinforcement learning

with lightweight knowledge distillation in mobile networks," *IEEE Journal on Selected Areas in Communications*, 2023.

[29] B. B. Gupta, K.-C. Li, V. C. Leung, K. E. Psannis, S. Yamaguchi *et al.*, "Blockchain-assisted secure fine-grained searchable encryption for a cloud-based healthcare cyber-physical system," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 12, pp. 1877–1890, 2021.

[30] G. N. Nguyen, N. H. Le Viet, M. Elhoseny, K. Shankar, B. Gupta, and A. A. Abd El-Latif, "Secure blockchain enabled cyber–physical systems in healthcare using deep belief network with resnet model," *Journal of parallel and distributed computing*, vol. 153, pp. 150–160, 2021.

[31] A. Raj and S. Prakash, "A privacy-preserving authentic healthcare monitoring system using blockchain," *International Journal of Software Science and Computational Intelligence (IJSSCI)*, vol. 14, no. 1, pp. 1–23, 2022.

[32] R. Ramanath, F. Liu, N. Sadeh, and N. A. Smith, "Unsupervised alignment of privacy policies using hidden markov models," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 605–610.

[33] K. Zhao, L. Yu, S. Zhou, J. Li, X. Luo, Y. F. A. Chiu, and Y. Liu, "A fine-grained chinese software privacy policy dataset for sequence labeling and regulation compliant identification," *arXiv preprint arXiv:2212.04357*, 2022.

[34] W. U. Ahmad, J. Chi, T. Le, T. Norton, Y. Tian, and K.-W. Chang, "Intent classification and slot filling for privacy policies," *arXiv preprint arXiv:2101.00123*, 2021.

[35] S. Zimmeck, P. Story, D. Smullen, A. Ravichander, Z. Wang, J. R. Reidenberg, N. C. Russell, and N. Sadeh, "Maps: Scaling privacy compliance analysis to a million apps," *Proc. Priv. Enhancing Tech.*, vol. 2019, p. 66, 2019.

[36] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell *et al.*, "The creation and analysis of a website privacy policy corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1330–1340.

[37] E. Poplavska, T. B. Norton, S. Wilson, and N. Sadeh, "From prescription to description: Mapping the gdpr to a privacy policy corpus annotation scheme," in *Legal Knowledge and Information Systems-JURIX 2020: 33rd Annual Conference*, 2020.

[38] H. Al-Khalifa, M. Mashaabi, G. Al-Yahya, and R. Alnashwan, "The saudi privacy policy dataset," *arXiv preprint arXiv:2304.02757*, 2023.

[39] M. Srinath, S. Wilson, and C. L. Giles, "Privacy at scale: Introducing the privaseer corpus of web privacy policies," *arXiv preprint arXiv:2004.11131*, 2020.

[40] L. Lebanoff and F. Liu, "Automatic detection of vague words and sentences in privacy policies," *arXiv preprint arXiv:1808.06219*, 2018.

[41] S. Arora, H. Hosseini, C. Utz, V. K. Bannihatti, T. Dhellemmes, A. Ravichander, P. Story, J. Mangat, R. Chen, M. Degeling *et al.*, "A tale of two regulatory regimes: Creation and analysis of a bilingual privacy policy corpus," in *LREC proceedings*, 2022.

[42] A. Ravichander, A. W. Black, S. Wilson, T. Norton, and N. Sadeh, "Question answering for privacy policies: Combining computational and legal perspectives," *arXiv preprint arXiv:1911.00841*, 2019.

[43] M. Barati and O. Rana, "Tracking gdpr compliance in cloud-based service delivery," *IEEE Transactions on Services Computing*, vol. 15, no. 3, pp. 1498–1511, 2020.

[44] N. B. Truong, K. Sun, G. M. Lee, and Y. Guo, "Gdpr-compliant personal data management: A blockchain-based solution," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1746–1761, 2019.

[45] L. Campanile, M. Iacono, F. Marulli, and M. Mastroianni, "Designing a gdpr compliant blockchain-based iov distributed information tracking system," *Information Processing & Management*, vol. 58, no. 3, p. 102511, 2021.

[46] H. Ahmad and G. S. Aujla, "Gdpr compliance verification through a user-centric blockchain approach in multi-cloud environment," *Computers and Electrical Engineering*, vol. 109, p. 108747, 2023.

[47] D. Torre, S. Abualhaija, M. Sabetzadeh, L. Briand, K. Baetens, P. Goes, and S. Forastier, "An ai-assisted approach for checking the completeness of privacy

policies against gdpr," in *2020 IEEE 28th International Requirements Engineering Conference (RE)*. IEEE, 2020, pp. 136–146.

[48] S. D. Gupta and T. Hahmann, "Oppo: An ontology for describing fine-grained data practices in privacy policies of online social networks," *arXiv preprint arXiv:2309.15971*, 2023.

[49] H. Cui, R. Trimananda, A. Markopoulou, and S. Jordan, "{PoliGraph}: Automated privacy policy analysis using knowledge graphs," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 1037–1054.

[50] S. Tokas, O. Owe, and T. Ramezanifarkhani, "Static checking of gdpr-related privacy compliance for object-oriented distributed systems," *Journal of Logical and Algebraic Methods in Programming*, vol. 125, p. 100733, 2022.

[51] K. Hjerppe, J. Ruohonen, and V. Leppänen, "Annotation-based static analysis for personal data protection," *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14*, pp. 343–358, 2020.

[52] Y. Ling, K. Wang, G. Bai, H. Wang, and J. S. Dong, "Are they toeing the line? diagnosing privacy compliance violations among browser extensions," in *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2022.

[53] S. Liu, F. Zhang, B. Zhao, R. Guo, T. Chen, and M. Zhang, "Appcorp: a corpus for android privacy policy document structure analysis," *Frontiers of Computer Science*, vol. 17, no. 3, p. 173320, 2023.

[54] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[55] F. Wang, H. Zhu, G. Srivastava, S. Li, M. R. Khosravi, and L. Qi, "Robust collaborative filtering recommendation with user-item-trust records," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 4, pp. 986–996, 2021.

[56] X. Zhou, W. Liang, I. Kevin, K. Wang, and L. T. Yang, "Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 171–178, 2020.

[57] L. Qi, W. Lin, X. Zhang, W. Dou, X. Xu, and J. Chen, "A correlation graph based approach for personalized and compatible web apis recommendation in mobile app development," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Jintao Chen is working towards a Ph.D. degree at the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. She received her B.S. degree in Internet of Things from Central South University. Her research interests include Service computing, Service Regulation, Process Mining, and Machine Learning.

Fan Wang is currently pursuing a Ph.D. degree at the College of Computer Science and Technology, Zhejiang University, Hangzhou, P.R. China. She received her master's degree from the School of Computer Science, Qufu Normal University, Rizhao, China, in 2021. Her research interests include big data analyses and recommendation system.

Shengye Pang is working towards a Ph.D. degree at the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. He received a B.S. degree in computer science from Tongji University, in 2015, and a Master's degree in computer science from Shanghai University, in 2019. His current research interests include Service Computing and Game Theory.

Mingshuai Chen received the Ph.D. degree in computer science from the Institute of Software, Chinese Academy of Sciences, Beijing, China, in 2019. He then worked as a postdoctoral researcher at the Department of Computer Science, RWTH Aachen University, Aachen, Germany. Since 2023, he joined the College of Computer Science and Technology, Zhejiang University, Hangzhou, China as an assistant professor. His main research interest lies in formal methods for emerging computing paradigms.

Meng Xi received his B.S. degree in computer science from Zhejiang University, China in 2017. He is now working towards a Ph.D. degree at the College of Computer Science and Technology, Zhejiang University, Hangzhou, China, and is supported by the China Scholarship Council for 1 year's study at the Nanyang Technological University, Singapore. He has been a recipient of the Best Paper Award of IEEE SMDS 2020. His current research interests include Service Computing, Data Science, and Machine Learning.

Tiancheng Zhao received his Ph.D. in Computer Science from Carnegie Mellon University (CMU), advised by Prof. Maxine Eskenazi. He received his B.S in Electrical Engineering from University of California, Los Angeles (UCLA) with Summa Cum Laude. He is currently a principal researcher at Binjiang Insititute of Zhejiang University and has published more than 30 papers in top journals and conferences. His current research interests include multimodal machine learning, natural language processing, and service computing.

Jianwei Yin received his Ph.D. degree in computer science from Zhejiang University (ZJU) in 2001. He was a Visiting Scholar with the Georgia Institute of Technology. He is currently a Full Professor with the College of Computer Science, ZJU. Up to now, he has published more than 100 papers in top international journals and conferences. His current research interests include service computing and business process management. He is an Associate Editor of the IEEE Transactions on Services Computing.