User Privacy Risk Analysis within Website Privacy Policies

Thi-Huong-Giang Vu
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
giangvth@soict.hust.edu.vn
https://orcid.org/0000-0002-2522-4241

Xuan-Bach Hoang
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi, Vietnam
bach.hx194481@sis.hust.edu.vn
https://orcid.org/0009-0003-1155-1828

2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR) | 979-8-3503-6843-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/MAPR63514.2024.10660854

***Abstract*—The websites' privacy policies reflect how their users' personal data is collected, processed, stored, and used. This paper proposes a risk analysis method to assist users in understanding and identifying potential privacy risks in these policies. The method consists of two phases. First, we propose a transfer learning process to refine an existing large language model (LLM) to enable it to understand and answer legal questions effectively. Second, we introduce a retrieval-augmented generation (RAG) technique applied to the refined LLM to analyze risks in privacy policies. Additionally, we define a checklist to assess the compliance of privacy policies with legal regulations. This checklist is mapped to 20 queries, which are used to build prompts for the refined LLM. This allows our refined LLM to provide explanatory answers to users regarding the overall compliance level of privacy policies with relevant regulations, stating the potential risks in these privacy policies. The transfer learning process has been conducted on two existing LLMs. Experiments were performed on 200 random legal questions to compare these fine-tuned LLMs with the originals, demonstrating the efficiency of our proposed method in terms of the cosine similarity index and the F1 score.**

***Index Terms*—Privacy policy, legal compliance, LLM, RAG, privacy risk analysis**

I. Introduction

The advancement of technology and the Internet, coupled with the explosion of online platforms and e-commerce, have posed new challenges in protecting users' privacy rights, leading to an increased interest in the domain. From the governmental perspective, 71% of countries in the world currently have legislation concerning data protection and privacy. In Vietnam, the right to privacy is recognized in Article 21 of the 2013 Constitution [1]. The 2015 Civil Code [2] and the 2015 Penal Code [3], amended in 2017 [4] have specified the regulations on individual privacy rights and remedies when violating privacy rights, in line with international regulations on human privacy rights. The protection of privacy in general, and the privacy of personal data in particular, continue to be regulated by various laws and decrees such as: the law on cyber security [5], the law on information technology [6], the law on protection of consumers' rights [7], the law on cyber information security [8], the decree on protection of personal data [9], and the decree on e-commerce [10]. However, there are still many flaws in the application and enforcement of these regulations.

From the perspective of online service providers, a privacy policy is a legal document that outlines the provider's commitments and legal obligations, which users usually must accept before using the service. A transparent privacy policy typically includes general information about the business, the processes of collecting, storing, processing, and using users' data gathered during service operation, the information protection measures, and users' rights to their personal data. Additionally, these policies provide information about data sharing with third parties and a commitment to comply with relevant legal regulations. Therefore, a privacy policy is crucial in creating a safe and trustworthy online platform, building the users' trust and satisfaction and reinforcing the provider's reputation. As personal data becomes a valuable resource in the digital age, not all providers stick lawfully and fairly to all articles contained in privacy policies. Some e-commerce websites use non-transparent practices to take advantage of users and violate current privacy regulations. Loopholes in privacy policies can put users in a position of lack of support or force them to take responsibility when legal issues arise, creating potential risks. This situation requires stricter monitoring of compliance with regulations to protect users' personal information in a comprehensive and fair manner.

From the users' perspective, understanding these policies is challenging and often tedious because they are written in complex legal language and contain many technical terms. Most privacy policies require users to have a college-level reading ability [11]. This could lead users to agree to terms in the privacy policy that they did not clearly understand, particularly regarding potential risks, such as providers not applying security measures to the data they collect, leading to data leaks, misuse and exposure.

In this context, it's crucial to facilitate users' understanding of providers' privacy policies, whether they comply with legal regulations or not, and to identify the inherent risks in these policies for users. Analyzing and evaluating these risks is extremely important to protect users' privacy and personal information. This enables users to recognize and proactively address potential risks, thereby taking appropriate

979-8-3503-6843-7/24/$31.00 ©2024 IEEE
Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on October 06,2024 at 00:24:03 UTC from IEEE Xplore. Restrictions apply.

preventive measures. Large language models (LLMs) are one of promising approaches to address the legal compliance problem. However, most of them [12], [13] face significant limitations when analyzing privacy policy documents. First, the context length they support is up to 4096 tokens, while the tokens needed to describe the context often exceed this limit multiple times. Second, hallucinations in their responses could lead to misunderstandings or inaccurate information during the analysis of privacy policies. Third, they may lack awareness of new events occurring after the training period, making it challenging to cite related legal articles accurately. Finally, retraining a LLM with billions of parameters is both time-consuming and costly.

This paper proposes a method for analyzing potential privacy risks in the provider's privacy policies. The method includes a LLM transfer learning process for enhancing their ability of providing detailed explanations to users' legal queries, and a legal compliance check list for identify risks. The enhanced LLMs provides answers that enable users to determine whether the provider's privacy policies comply with legal regulations and accurately reflect the inherent risks associated with these policies. Our paper has three main contributions. First, our method improved large language models capable of analyzing and answering privacy-related questions. Second, our method established a large text analysis mechanism with input constraints for large language models. Finally, our method enhanced the accuracy of privacy responses through the use of retrieval-augmented generation (RAG).

The article is structured as follows. Section II summarizes the related research and applications. Section III describes the transfer learning process for enhancing an existing LLM into a legal LLM. Section IV presents the privacy policy analysis process using the enhanced model. Section V details the experimentation process and the results obtained from the experiments. Finally, section VI summarizes the research findings, provides conclusions, and suggests directions for future development.

II. Related Works

Related studies are divided into two groups: studies on determining the legal compliance of privacy policies and studies on legal questions and answers.

A. Determining the legal compliance of privacy policies

Most studies have primarily focused on the laws of developed countries, such as the European General Data Protection Regulation (GDPR) [14] and the Singapore Personal Data Protection Act (PDPA) [15], rather than the laws of developing countries. In [16], BERT model is used to assess privacy policy compliance with GDPR and PDPA. The level of compliance (ranging from 0 to 5) is determined by classifying and labeling text segments in privacy policies and then comparing cosine similarity with GDPR and PDPA documents.

Beside, several compliance-checking tools have been proposed for end-users and providers. The GDPR Compliance Framework [17] provides a list of questions to assist website providers in building GDPR-compliant privacy policies. PrivacyCheck v3 [18] is a machine learning-based tool that allows users to summarize and score privacy policies by answering questions related to user control and GDPR compliance. In [19], a NLP-based framework for evaluating legal compliance in privacy policies has been presented, highlighting the impact of GDPR on the content of these policies.

B. Legal questions and answers

Many efforts have been invested in building legal chatbots. In [20], a cooperative-legal question-answering LLM-based chatbot has been proposed to address a set of legal questions about Palestinian cooperatives. In [21], an automated chatbot for predicting legal case outcomes has been developed. These studies have highlighted the potential of using LLMs to help users access and enhance their understanding of legal documents.

III. Transfer Learning Process for Building Legal LLM

As illustrated in Fig. 1, the transfer learning process of building a legal LLM model that supports risk analysis consists of 2 steps: incremental pre-training and supervised fine tuning.

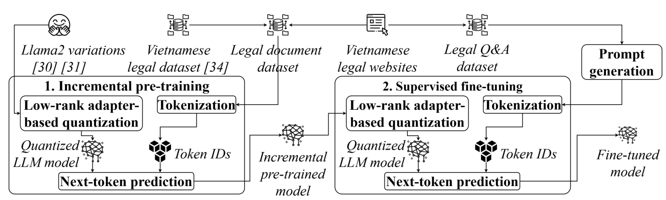

Fig. 1. Transfer learning process

A. Incremental pre-training

A pretrained LLM is selected to incremental pre-training on our legal document dataset, using a self-supervised method named next-token prediction to enhance its knowledge of Vietnamese legal issues. During the model's incremental pre-training process, a low-rank adapters-based quantization method is used to reduce the memory cost and speed-up the pre-training.

1) Next-token prediction: The next-token prediction technique is used to incremental pretrain the quantized model on the legal document dataset. For that, some legal articles from this dataset are sampled and used to train the model to predict the token that comes after each token within the sampled documents.

2) Quantization: The quantization allows reducing the size of original pretrained weights within this LLM to 4-bit precision. The low-rank adapters are used to capture gradient updates on the LLM's weights during pre-training. Each adapter corresponds to a matrix with rank lower than the minimum number of its rows and its columns, and its elements are stored with 16-bit precision.

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on October 06,2024 at 00:24:03 UTC from IEEE Xplore. Restrictions apply.

The output of the incremental pre-training step is a model with enhanced knowledge of Vietnamese legal issues (legal LLM).

B. Supervised fine-tuning the legal LLM

The legal LLM are fine-tuned to effectively answer legal questions. For this purpose, the supervised fine-tuning dataset is built based on question-and-answer pairs focused on privacy issues. These pairs, containing keywords such as personal, privacy, data, e-commerce, and information security, are crawled from legal websites. Next, pairs that exceed the selected LLM's input limit (4096 tokens) are removed to prevent halting the fine-tuning process due to VRAM resource overflow. Before training, these pairs are prompted following the standard prompt of the selected LLM. The next-token prediction method is also used for supervised fine-tuning. Rather than predicting from the first token of a question, supervised fine-tuning allows predicting the next token from the last token x_i for a question with a token length of i.

IV. Privacy Policy Analysis

The proposed privacy policy analysis process is divided into 2 steps: input enrichment and user privacy risk identification.

A. Input enrichment

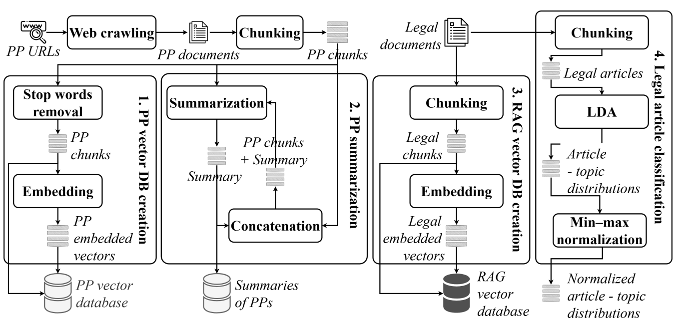

Fig. 2. Input enrichment process

As illustrated in Fig. 2, the input enrichment process consists of four steps. The first and second steps aim to extract features from privacy policy documents by creating a privacy policy vector database and summarizing the policies. The third step aims to create a RAG vector database for retrieving Vietnamese legal context during the user privacy risk identification phase. The last step aims to classify articles in legal documents into topics to align them with the user privacy risk checklist.

1) Privacy policy vector database creation: Entire privacy policy documents are automatically crawled from websites' URLs and then parsed. Stop words are removed from these documents, which are then divided into chunks using recursive strategy. The resulting chunks are encoded as vectors and stored locally in a vector database.

2) Privacy policies summarization: Summaries of privacy policies are produced to prevent information loss in the privacy policy chunks. To achieve better accuracy in summaries, rather than focusing on performance, we apply the refine summarization method combined with Google Gemini model [22]. The refine summarization method ensures continuity in context by looping the summarization process as follows. Initially, the privacy policy document is segmented into chunks of appropriate size for the Gemini model context. The first chunk is directly fed into the Gemini model to generate a summary. This summary is then combined with the second chunk for further summarization. This process is repeated until all chunks are processed.

3) RAG vector database creation: Crawled legal documents are chunked without removing stop words and encoded as RAG vectors. These vectors are then stored in a database.

The outputs of the input enrichment step include summaries of privacy policies, privacy policy vectors, and RAG vectors. These outputs are then passed into the LLM for analysis.

4) Legal articles classification: First, Vietnamese privacy regulations [6], [9], [10] are chunked into articles. Then, a probabilistic model called Latent Dirichlet Allocation (LDA) [23] is used to discover the underlying topics within these articles. This model assigns distribution probabilities of topics to each article. Next, the min-max normalization technique is applied to ensure that all topic distribution probabilities for each article are scaled to a range between 0 and 1. Finally, a threshold of 0.9 is applied to determine the dominant topics in an article, and then to decide which topics an article belongs to.

B. User privacy risk identification

The user privacy risk identification process is illustrated in Fig. 3. We define a checklist to identify risks. Its criteria are mapped to 20 queries, which are used to build prompt patterns for our fine-tuned LLM model. The LLM's answers are then coupled with the checklist compliance score to formulate responses about potential risks in the privacy policy, clearly stating these risks.

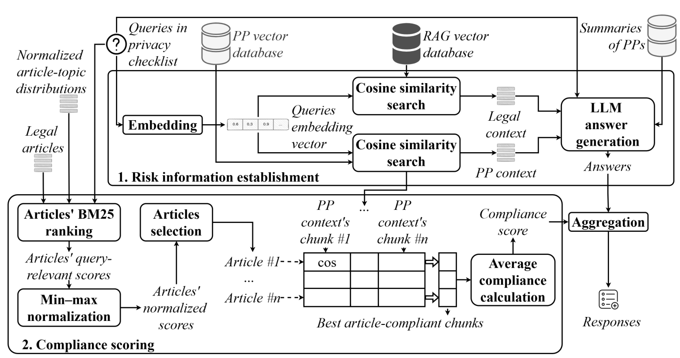

Fig. 3. User privacy risk identification

1) Privacy risks checklist definition: The checklist is developed based on Vietnamese privacy regulations and comprises of four criteria: privacy rights, lawful basis and transparency,

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on October 06,2024 at 00:24:03 UTC from IEEE Xplore. Restrictions apply.

data security, and accountability and governance. The 20 queries belonging to these criteria are illustrated in Tab. I.

TABLE I
CHECK LIST OF USER PRIVACY RISKS

#	Criteria
Lawful basis and transparency: Does the privacy policy clearly state something?	
1	How the provider collects, processes, and uses users' personal information
2	Activities related to the processing of users' personal data
3	Who, including individuals, organizations, or third parties, has access to users' personal information?
4	Provisions related to children and special categories of personal data
5	The types of personal data collected by the provider and when this data is collected
6	The purposes for which the provider collects personal information
Privacy rights: Does this privacy policy specify that users have the right to do something?	
7	Request and receive support related to their personal information, and to know the duration of data storage
8	Check, update, modify, or request the correction of their personal information
9	Delete or request the deletion of their personal data
10	Request the restriction of the processing of their personal data
11	Withdraw their consent
12	Request access to their personal data
13	Object to the processing of their personal data to prevent or restrict the disclosure of personal data or its use for advertising and marketing purposes
14	Consent or refuse consent to the processing of their personal data
15	Request compensation for damages in accordance with the law in case of violations of the protection of their personal data
16	Request businesses and individuals to check, modify, update, delete, transfer, or stop transferring their information to third parties
Data security: Does this privacy policy clearly ensure something?	
17	The protection and security measures during processing, including protection against violations of personal data protection regulations and prevention of loss, destruction, or damage due to incidents (e.g., encryption)?
18	Have procedures in place for notifying authorities and data subjects in the event of a personal data protection violation?
Accountability and governance: Does this privacy policy provide something?	
19	Information such as the name, registered address, and contact phone number of the merchant, organization, or individual owning the website?
20	A commitment to complying with Vietnamese regulations on personal data protection?

2) Risk information establishment: First, a query is selected from the privacy risks checklist. Second, the query is transformed as a vector. The vector representing the query is compared with the vectors of chunks stored in the privacy policy vector database to find the top k chunks with the greatest similarity to the query. Thirds, the query vector is compared with the vectors of chunks stored in the RAG vector database to find the top k chunks with the greatest similarity to the query. The prompt is built based on the results of second and third steps, the summary of privacy policy, and the query itself. This prompt is post into fine-tuned models for analyzing the privacy policy document.

3) Compliance scoring: Firts, the relevance of legal articles to a given query in the checklist is evaluated by using the BM25 ranking function [24]. Each article are scored based on the frequency of the query's terms within the article (TF), their occurrence across all legal documents (IDF), and considering the article length. The BM25 scores are then normalized to the range [0, 1] using the min-max normalization. Articles with a normalized BM25 score above 0.5 are selected for further evaluation. Next, the compliance score of a privacy policy is calculated based on the cosine similarity of its chunks with the selected articles. For each chunk of the privacy policy, the article with the highest cosine similarity is identified. The compliance score of the privacy policy is determined by averaging these highest cosine similarity values across its chunks. The combination of the LLM's answer with the privacy policy's legal compliance score allows identifying potential risks in the privacy policy and also its legal compliance.

V. EXPERIMENTATION

Our proposed methods presented in Section III and Section IV is implemented by using some library such as QLoRA [25], Transformers [26], Scrapy [27], Selenium [28], and LangChain [29].

We chose 2 pre-trained variations of LLama2 with 7 billion parameters to pre-training and fine-tuning. The first is the vbd-llama2-7B-50b-chat model of the LR-AI-Labs [30], hereafter referred to as LLM1. The second is the vietnamese-llama2-7b-120GB model of BKAI [31], hereafter referred to as LLM2.

For training, we use the free configuration of the Kaggle platform [32] with 16GB VRAM, P100 GPU, 29GB RAM, 73GB disk and 12-hour session limit. Such configuration for allows training up to 33,000 samples per session. This setup allows for training up to 33,000 samples per session. However, since our dataset exceeds this capacity, we divide it into smaller subsets, each containing a maximum of 31,000 samples. To track training results across these subsets, we integrated a logging tool called Wandb [33]. This tool ensures that metrics such as loss, evaluation scores, and other relevant information are logged and can be continued in future sessions. Additionally, we use this tool to save model checkpoints during training, ensuring that the training process can resume if the session is interrupted due to exceeding the session time limit.

A. Data preparation

The dataset is divided into two types: legal document dataset and legal Q&A dataset.

Legal document dataset: is about 7.94 GB in plain text format, including 2 parts. Part 1 is a corpus of Vietnamese legal documents published before 2021, sourced from Kaggle [34]. Part 2 consists of recently updated Vietnamese legal documents related to user privacy and relevant press articles, regularly crawled from various sources by our crawling module. After preparation, our training dataset comprises 176,578 samples from Part 1 and 7,265 samples from the Part 2. Our test set consists of 999 samples from Part 1 and 200 samples from Part 2.

Legal Q&A dataset: contains total 24,797 pairs of questions and answers related to user privacy. This includes the following fields: $original\_url$ (the link to the original article),

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on October 06,2024 at 00:24:03 UTC from IEEE Xplore. Restrictions apply.

$Q\&A$ (the content of the question and its corresponding answer, a.k.a. desired answer), and $length$ (the total length of the question and answer). The result after filtering the data is a training set with 10,429 samples and a test set with 1,159 samples.

B. Training parameters and results

1) Parameters: During the model training process, we use a series of parameters as illustrated in Table II. These parameters are carefully selected to ensure the optimal performance of the model during training. lora_r is related to the number of trainable parameters with LoRA. lora_dropout prevents overfitting in training process. epochs is the number of times the model will train on the entire dataset. batch_size is the number of data samples used in one gradient update.

TABLE II
PRE-TRAINING AND FINE-TUNING PARAMETERS

Step	lora_r	lora_dropout	epochs	batch_size
Pre-training	16	0.1	1	1
Fine-tuning	16	0.1	3	2

2) Results:

a) Incremental pre-training results: We incrementally trained both LLM1 and LLM2 for 90 hours, using about 183,843 samples of the legal document dataset. However, the experimental results of the incrementally trained models were worse than those of the original models. There could be two reasons for this. First, LLM1 and LLM2 were previously trained on Vietnamese legal datasets, so the incremental pre-training did not yield the expected improvements. Second, during training, our prompting process was not as effective, causing the models not being able to complete the prediction process as expected.

b) Supervised fine-tuning results: We conducted this step on the original LLM1 and the original LLM2 using the entire legal Q&A dataset referenced in Section V-A. The training duration was approximately 50 hours. The results demonstrated a notable enhancement in quality for the legal Q&A tasks.

C. LLMs evaluation

The experiment was conducted on an evaluation set consisting of 200 samples of random legal questions. Table III presents the cosine similarity and F1 scores of the original LLMs and the fine-tuned LLMs.

Regarding cosine similarity, the experimental results demonstrate that our fine-tuned models achieved notably high scores, with the fine-tuned LLM1 model achieving the highest score. The LLM1's cosine score is higher than LLM2's cosine score, both before and after fine-tuning. This can be explained by the original LLM1 being specifically trained for question-answering tasks, whereas the original LLM2 has not been fine-tuned on specific tasks. After fine-tuning, both LLM1 and LLM2 showed significant improvements in their performance on the legal question-answering task, with LLM1 performing slightly better overall.

The F1 score is used to measure the lexical similarity of the LLM's answers compared to the desired answers. Accordingly, the score of the trained LLM2 has improved compared to the original models, demonstrating the effectiveness of the training process.

TABLE III
LLMS' EVALUATION

Model name	Cosine similarity index	F1 score
Original LLM1 [30]	0.853577	0.195341
Original LLM2 [31]	0.486425	0.053904
Fine-tuned LLM1	0.877830	0.210391
Fine-tuned LLM2	0.868871	0.189713

D. Results of risks analysis

We conducted this experiment with our fine-tuned LLM1 to identify user privacy risks in the privacy and security policies of 100 Vietnamese websites. Our model generated 20 responses for each policy, based on our privacy risks checklist, with each response including an answer and a compliance score from 0 (non-compliant) to 1 (fully compliant). The quality of the analysis of potential privacy risks is evaluated using the criteria in Table IV.

TABLE IV
ANSWERS' EVALUATION CRITERIA

Criteria	Description
Relevance	Directly address the question
Detail	Identify potential risks in the privacy policy regarding the question
Accuracy	Reflect the content of the privacy policy
Analysis	Provide further information about the identified risks
Solution	Propose preventive measures or guidance for users to manage the identified risks

Below are two examples of the analysis of the privacy policy of Thanh Niên newspaper [35] by asking our fine-tuned LLM1 two questions #8 and #17.

1) Example 1: Privacy rights analysis with Question #8

Answer: Based on the information provided in the privacy policy of Thanh Nien newspaper, users can easily correct or update inaccurate or incomplete information by contacting Thanh Nien newspaper. However, there are some risks associated with correcting or updating this information. One risk is that the corrected or updated information might still be inaccurate or incomplete, leading to misunderstandings or security issues. Another risk is that users might provide inaccurate or incomplete personal information, resulting in the inaccuracy or incompleteness of their records. Additionally, there is a risk that the corrected or updated information could be used for non-profit purposes or worse purposes without the user's consent. Overall, while users can correct or update their personal information, it is important to consider the potential risks and ensure that the corrected or updated information is accurate and complete.

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on October 06,2024 at 00:24:03 UTC from IEEE Xplore. Restrictions apply.

Compliance score: 0.38.

This answer correctly identifies a legal non-compliance with the policy. The risk that arises from the lack of a data breach notification procedure is also noted.

2) Example 2: Data security analysis with Question #17

Answer: The policy also does not disclose measures for protection, security during processing, including safeguarding against violations of regulations regarding personal data protection and prevention of loss, destruction, or damage due to unforeseen incidents. Risks for users in this regard may include: The provider may be subject to cyber attacks or hacking, resulting in the exposure of users' personal information.

Compliance score: 0.55.

In this answer, the legal non-compliance with the policy is clearly stated, specifically noting the absence of details regarding protective and security measures. The risk of a cyber attack is also highlighted.

By asking the same questions (#8 and #17) to ChatGPT 3.5 and Gemini, and then asking them to compare answers according to the criteria in Table IV, we obtain the answers' ranking shown in Table V.

TABLE V
COMPARISON OF ANSWERS FROM DIFFERENT LLMS

		Ranking from Gemini	Ranking from ChatGPT 3.5
Question #8	Fine-tuned LLM1	1	1
	Chat GPT 3.5	3	3
	Gemini	2	2
Question #17	Fine-tuned LLM1	1	3
	Chat GPT 3.5	3	2
	Gemini	2	1

VI. CONCLUSION

Knowledge about a website's privacy policies enables users to recognize and proactively address potential privacy risks. To enhance users' efficiency in analyzing privacy policy documents, our study introduces a solution that informs users about the compliance of privacy policies with relevant regulations and highlights potential risks within these policies. This solution is based on incrementally training a pre-trained LLM with updated legal knowledge, fine-tuning it to answer legal questions, and then defining a checklist comprising 20 queries to prompt the LLM about potential risks. The legal compliance score of privacy policies are also calculated. Future work includes incorporating RAG to create Q&A datasets with context and developing a machine learning-based method for scoring the risks.

REFERENCES

[1] "Constitution of the Socialist Republic of Vietnam," 2013.
[2] "Civil Code," 91/2015/QH13, 2015.
[3] "Criminal Code," 100/2015/QH13, 2015.
[4] "Law on Amendments to the Criminal Code No. 100/2015/QH13," 12/2017/QH14, 2017.
[5] "Law on Cyber Security," 24/2018/QH14, 2018.
[6] "Law on Information Technology," 67/2006/QH11, 2006.
[7] "Law on Protection of Consumers Rights," 19/2023/QH15, 2023.
[8] "Law on Cyber Information Security," 86/2015/QH13, 2015.
[9] "Decree on Protection of Personal Data," 13/2023/ND-CP, 2023.
[10] "Decree on E-commerce," 13/2023/ND-CP, 2013.
[11] M. A. Graber, D. M. D'Alessandro, and J. Johnson-West, "Reading level of privacy policies on Internet health Web sites." *The Journal of family practice*, vol. 51 7, pp. 642–5, 2002.
[12] "Llama 2: open source, free for research and commercial use," Meta. [Online]. Available: https://llama.meta.com/llama2/
[13] M. AI, "Mistral 7B the best 7B model to date, Apache 2.0," https://mistral.ai/news/announcing-mistral-7b/.
[14] E. Union, "General data protection regulation (GDPR)," 2016.
[15] "Singapore Personal Data Protection Act," 2012.
[16] A. Qamar, T. Javed, and M. O. Beg, "Detecting compliance of privacy policies with data protection laws," 2021, arXiv:2102.12362 [cs]. [Online]. Available: https://arxiv.org/abs/2102.12362
[17] M. Ferreira, T. Brito, J. F. Santos, and N. Santos, "RuleKeeper: GDPR-aware personal data compliance for web frameworks," in *2023 IEEE Symposium on Security and Privacy (SP)*, 2023, pp. 2817–2834.
[18] R. Nokhbeh Zaeem, A. Ahbab, J. Bestor, H. H. Djadi, S. Kharel, V. Lai, N. Wang, and K. S. Barber, "Privacycheck v3: Empowering users with higher-level understanding of privacy policies," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2022, p. 1593–1596.
[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)*, 2022, pp. 1–9.
[20] R. Qasem, B. Tantour, and M. Maree, "Towards the exploitation of LLM-based chatbot for providing legal support to Palestinian cooperatives," Jun. 2023, arXiv:2306.05827 [cs]. [Online]. Available: http://arxiv.org/abs/2306.05827
[21] S. Alam, R. Pande, M. S. Ayub, and M. A. Khan, "Towards developing an automated chatbot for predicting legal case outcomes: A deep learning approach," in *Intelligent Information and Database Systems*, N. T. Nguyen, S. Boonsang, H. Fujita, B. Hnatkowska, T.-P. Hong, K. Pasupa, and A. Selamat, Eds. Singapore: Springer Nature Singapore, 2023, pp. 163–174.
[22] "Build with the Google Gemini API and Gemma open models," Google AI for Developers. [Online]. Available: https://ai.google.dev/gemini-api/docs/models/gemini?hl=en
[23] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 169–15 211, 2019.
[24] G. Amati, "BM25," Boston, MA, pp. 257–260, 2009. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_921
[25] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," 2023.
[26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2020.
[27] "Scrapy 2.11 documentation," Zyte. [Online]. Available: https://docs.scrapy.org/en/latest/
[28] "selenium 4.22.0," PyPI. [Online]. Available: https://pypi.org/project/selenium/
[29] "LangChain 0.2," LangChain. [Online]. Available: https://python.langchain.com/v0.2/docs/introduction/
[30] LR-AI-Labs, "vbd-llama2-7B-50b-chat," Hugging Face, 2023. [Online]. Available: https://huggingface.co/LR-AI-Labs/vbd-llama2-7B-50b-chat
[31] BKAI, "vietnamese-llama2-7b-120gb," Hugging Face, 2023. [Online]. Available: https://huggingface.co/bkai-foundation-models/vietnamese-llama2-7b-120GB
[32] "Kaggle: Your machine learning and data science community," Kaggle. [Online]. Available: https://www.kaggle.com
[33] "Wandb," Weights & Biases. [Online]. Available: https://github.com/wandb/wandb
[34] T. Q. Buu, "Vietnamese legal dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/quangbut/vietnamese-legal
[35] "Security privacy," Thanh Nien. [Online]. Available: https://thanhnien.vn/stories/chinh-sach-bao-mat

Authorized licensed use limited to: Univ of Calif Berkeley. Downloaded on October 06,2024 at 00:24:03 UTC from IEEE Xplore. Restrictions apply.