Using LLMs for Automated Privacy Policy Analysis: Prompt Engineering, Fine-Tuning and Explainability

Anonymous ACL submission

Abstract

Privacy policies are widely used by digital services and often required for legal purposes. Many machine learning based classifiers have been developed to automate detection of different concepts in a given privacy policy, which can help facilitate other automated tasks such as producing a more reader-friendly summary and detecting legal compliance issues. Despite the successful applications of large language models (LLMs) to many NLP tasks in various domains, there is very little work studying the use of LLMs for automated privacy policy analysis, therefore, if and how LLMs can help automate privacy policy analysis remains under-explored. To fill this research gap, we conducted a comprehensive evaluation of LLM-based privacy policy concept classifiers, employing both prompt engineering and LoRA (low-rank adaptation) fine-tuning, on four state-of-the-art (SOTA) privacy policy corpora and taxonomies. Our experimental results demonstrated that combining prompt engineering and fine-tuning can make LLM-based classifiers outperform other SOTA methods, *significantly* and *consistently* across privacy policy corpora/taxonomies and concepts. Furthermore, we evaluated the explainability of the LLM-based classifiers using three metrics: completeness, logicality, and comprehensibility. For all three metrics, a score exceeding 91.1% was observed in our evaluation, indicating that LLMs are not only useful to improve the classification performance, but also to enhance the explainability of detection results.

1 Introduction

In the digital age, the exponential growth of online services and applications has precipitated substantial concerns pertaining to user privacy protection. Some services or applications tend to excessively collect or utilize users' personal information, posing threats to privacy security. Privacy policies, serving as formal legal documents that delineate organizational data practices, constitute a critical mechanism for informing users about the collection, processing, storage and sharing of their personal data. The examination of privacy policies is of paramount importance for comprehending personal data processing mechanisms and evaluating organizational compliance with established privacy regulations such as the EU and the UK's GDPR (General Data Protection Regulation) (Voigt and Von dem Bussche, 2017) and the USA's CCPA (California Consumer Privacy Act) (California State Legislature, USA, 2018). However, privacy policies are often complex, filled with technical terms, making comprehension challenging. Past research (Ibdah et al., 2021) has revealed that many users encountered difficulties in understanding the content of privacy policies. Therefore, analyzing privacy policies in a way that facilitates user understanding and comprehension holds significant practical value. Due to the increasing number of online services and applications and the iterative nature of privacy policies, manual analysis becomes unsustainable, making machine learning based automated privacy policy analysis a meaningful research direction.

Large language models (LLMs) have demonstrated the state-of-the-art performance in various natural language processing (NLP) benchmarks, showcasing a remarkable potential in practical applications such as text generation, dialogue processing, and knowledge question-answering (Chang et al., 2024). It is highly likely that LLMs will perform well in analyzing privacy policies written in natural language, as their capabilities can be effectively leveraged given the complex nature of these policies. Although many researchers have proposed machine learning based classifiers for automated privacy policy analysis, to the best of our knowledge, except the limited work by Goknil et al. (2024) on exploring the use of prompt engineering LLMs for this purpose, the potential of LLMs

remains largely unexplored.

In this paper, we report our comprehensive evaluation of utilizing both prompt engineering and LoRA (low-rank adaptation) fine-tuning to develop LLM-based privacy policy concept classifiers. We conducted experiments across four state-of-the-art (SOTA) privacy policy corpora/taxonomies and several mainstream LLMs, exploring the effects of different factors such as temperature and model size on the model performance. Explainability refers to the ability to explain or present the behavior of AI models in human-understandable terms (Zhao et al., 2024). In addition to assessing the detection performance of LLM-based privacy policy concept classifiers, we also studied how to use LLMs to explain the detection results using customized prompts for different concepts. We evaluated the model's explainability across three metrics: completeness, logicality, and comprehensibility. Our key contributions are as follows:

1) We conducted a systematic evaluation on how to use LLMs to conduct automated privacy policy analysis, which, to the best of our knowledge, represents the first comprehensive study of this kind. By leveraging both prompt engineering and LoRA fine-tuning, we managed to use LLMs to produce new privacy policy concept classifiers that can outperform other SOTA classifiers *significantly* and *consistently* across three mainstream open-source LLMs and four SOTA privacy policy corpora/taxonomies.

2) We systematically investigated the potential of using LLMs to explain detection results of LLM-based privacy policy concept classifiers. Based on the above-mentioned three metrics, our human-based assessment results demonstrate that LLMs can generate meaningful explanations with high satisfaction (a score exceeding 91.1% observed for all three metrics), although there are some shortcomings in logicality.

The remainder of this paper is structured as follows. Section 2 presents related work. Section 3 details our approach. Section 4 outlines the experiment setup and results. Section 5 explores the explainability of LLMs in privacy policy analysis. The last two sections conclude this paper and discuss limitations of the work, respectively.

2 Related Work

2.1 Privacy Policy Corpora

Annotated privacy policy datasets are crucial for the training and evaluation of machine learning models. A common annotation involves segmenting privacy policies and classifying these segments based on taxonomies derived from legal standards or real-world privacy policies. As the first and the most widely used privacy policy dataset, OPP-115 (Wilson et al., 2016) provides fine-grained annotations at the paragraph level. It encompasses 115 privacy policies from online services, with 3,792 paragraphs categorized into 12 privacy policy conceptual categories (which forms a mini-taxonomy). Each paragraph was independently annotated by three legal experts and assigned to one or multiple privacy policy concepts. Three more recent datasets were released in 2024. Among them, Tang et al. (2024) introduced GoPPC-150, a dataset featuring paragraph-level annotations and a more comprehensive taxonomy tailored to GDPR requirements. Two other new datasets, CAPP-130 (Zhu et al., 2023) and APPCP-100 (Zhang et al., 2024), focus on Chinese privacy policies, offering support for research on privacy policy analysis in a multilingual context.

2.2 Automated Privacy Policy Analysis

Automated privacy policy analysis encompasses various tasks such as concept classification (Srinath et al., 2021; Mousavi Nejad et al., 2020; Tang et al., 2024), summary generation (Zhu et al., 2023), question answering (Harkous et al., 2018), and the annotation of key information like opt-out options (Bannihatti Kumar et al., 2020). Among these, concept classification in privacy policies has been more extensively studied. It involves segmenting privacy policies and labeling each segment based on a taxonomy covering relevant concepts. This approach can facilitate readers with a quick understanding of key conceptual points in different parts of the privacy policy. In addition, the coverage of concepts serves as an important criterion for assessing a privacy policy's legal compliance against a given data protection law.

Some researchers (Torre et al., 2020; Mousavi Nejad et al., 2020; Mustapha et al., 2020; Srinath et al., 2021; Tang et al., 2024) employed NLP approaches to automatically analyze the content of a given privacy policy and evaluated it on privacy policy corpora, establishing

a stable baseline. Some others (Xiang et al., 2023; Cejas et al., 2024) adopted semantic role based approaches to do large-scale privacy policy completeness violation studies. However, there has been very little research on automated privacy policy analysis based on LLMs. The only past study we are aware of was done by Goknil et al. (2024), who looked at using prompt engineering LLMs for this purpose only.

2.3 Large Language Models

Large language models (LLMs), like OpenAI's GPT series (Radford et al., 2018) and Meta's Llama series (Touvron et al., 2023), possess immense parameter sizes and learning capabilities. A notable capability of LLMs is their rich contextual learning ability (Brown et al., 2020). Through carefully designed prompts, such as detailed task-specific instructions or a few illustrative examples, researchers can effectively guide models to generate targeted outputs. Many prompt engineering methods for LLMs have been developed in the past a few years (Schulhoff et al., 2024), e.g., Wei et al. (2022c) introduced the chain-of-thought (CoT) approach, which decomposes complex problems into intermediate reasoning steps, helping LLMs generate more logical and coherent responses. In addition to prompt engineering, which is more in the domain of zero- or few-shot training, fine-tuning is another effective way to improve LLMs' abilities of solving new tasks (Wei et al., 2022a). However, the time complexity and costs of full-parameter fine-tuning can be exceedingly high due to the huge number of parameters in LLMs. To mitigate this issue, more efficient fine-tuning methods have been extensively developed, such as adapter tuning, prefix tuning, prompt tuning and LORA (Ding et al., 2023; Li and Liang, 2021; Lester et al., 2021; Hu et al., 2022).

3 Methodology

3.1 Problem Formulation

The problem can be defined as a multi-class multi-label classification task of assigning a segment in a given privacy policy one or more concepts defined in a relevant taxonomy. Among all privacy policy taxonomies, the one supporting the privacy policy corpus GoPPC-150 (Tang et al., 2024) is the most advance and the first multi-level one, with fine-grained privacy policy concepts especially those related to the GDPR. A partial hierarchy of the GoPPC-150 taxonomy is illustrated as a directed acyclic graph (DAG) in Figure 1.

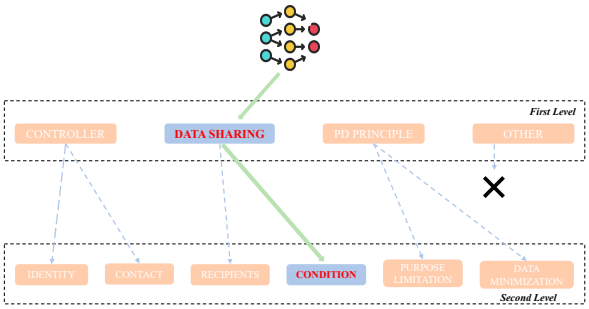

Figure 1: A partial hierarchy of GoPPC-150.

To elucidate the process of privacy policy concept classification, let us consider an illustrative example. Given a privacy policy segment X, the classifier $H(\cdot)$ produces a label Y_1 indicating one of the first-level nodes of the taxonomy, denoted as $Y_1 = H(X)$. A special value of Y_1 is 'OTHER', indicating that X does not match any first-level nodes in the taxonomy. If Y_1 does refer to a leaf node, such as 'DATA SHARING' the partial taxonomy in Figure 1, a subsequent classification task will proceed to determine the associated second-level node following a similar process, denoted as $Y_2 = H(X, Y_1)$, where Y_2 refers to the produced second-level node. The process can continue until a leaf node is reached, although for GoPPC-150 only the first two levels have sufficient data so the process will stop at the second level. The final classification result of X is therefore a cascaded code denoted by $Y_1.Y_2$, e.g., 'DATA SHARING.CONDITION'. Note that X may be labeled multiple concepts so more than one final classification code could be produced.

3.2 Design Prompts

We explored applying prompt engineering to privacy policy analysis. Given that the OPP-115 privacy policy corpus represents the first published and most widely used dataset in this domain, we utilized it as the benchmark to assess the effectiveness of various prompt designs. To comprehensively assess the impact of different prompt engineering techniques, such as few-shot and CoT, we designed five different prompts to elicit related concepts from privacy policy segments.

Each prompt was designed to provide different levels of guidance and context to LLMs. Figure 4 in Appendix A shows greater details of the five prompts.

- Prompt 1 simply describes the task without providing any additional explanations or examples.
- Prompt 2 extends Prompt 1 by including detailed explanations of all the 12 concepts defined in OPP-115.
- Prompts 3 and 4, in addition to providing category explanations, are designed for few-shot learning with one (Prompt 3) or two (Prompt 4) examples for each category.
- Prompt 5 introduces CoT as a structured reasoning process to guide LLMs through a step-by-step approach.

3.3 Fine-Tuning

We propose a method to streamline the adaptation of LLMs to hierarchical classification tasks, using the multi-level corpus GoPPC-150 as the benchmark dataset. The fine-tuning process involves two distinct tasks: predicting first-level nodes based on segment content, and subsequently predicting second-level nodes based on both segment content and the predicted first-level nodes. Figure 2 shows the two-leveled fine-tuning process. The process is progressive, and LLMs acquire two-stage prediction capabilities through this process.

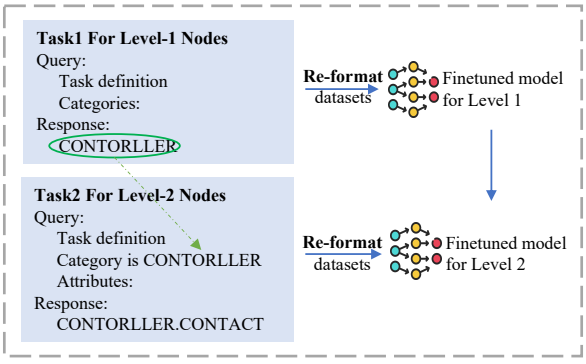

Figure 2: Two-leveled fine-tuning process.

4 Experiments and Results

4.1 Setup

Corpora. Different corpora employ different concept taxonomies. These taxonomies differ in both granularity, multi-level taxonomies being more detailed than single-level ones, and construction standard, with some tailored to specific regulations such as the GDPR and others being more general. OPP-115 uses a simple single-level taxonomy to broadly categorize concepts. The taxonomy it employs was defined based on real-world privacy policies by law experts over multiple iterations. In contrast, corpora like GoPPC-150 and APPCP-100 employ a multi-level taxonomy that is specifically designed to align with the hierarchical nature of privacy policies and a data protection regulatory framework, thereby offering greater relevance and applicability to real-world scenarios. For a comprehensive study of diverse taxonomies, we selected OPP-115 and GoPPC-150, complemented by their Chinese counterparts, CAPP-130 and APPCP-100, which adhere to a single-level and a multi-level taxonomy, respectively.

The OPP-115 corpus consists of 10 concept categories, with the final category 'OTHER' further subdivided into three distinct categories: 'Introductory/Generic', 'Privacy Contact Information', and 'Practice Not Covered'. We considered all the 12 categories as a single-level taxonomy for our experiments. The GoPPC-150 corpus has nodes at three levels, but only 14 first-level and 21 second-level nodes have sufficient data, which were used for our experiments. The CAPP-130 corpus classifies privacy policy segments by three aspects: importance, risk, and topic classification. This paper focuses on the topic classification task (with 11 topical categories) that is aligned with the classification of concepts. For APPCP-100, the nodes that appear infrequently are filtered and 13 first-level, 25 second-level, and 16 third-level nodes were selected for our experiments.

Models. To ensure that our experimental results align with the latest advancements of LLMs, we selected three widely recognized open-source models for our experiments. These include Llama developed by Meta (Touvron et al., 2023), the Qwen series developed by Alibaba (Bai et al., 2023), and ChatGLM developed by the Tsinghua University (GLM et al., 2024). We also conducted limited experiments using GPT, the most studied closed-source LLM from Open AI (Radford et al., 2018). These models were chosen because of their strong performance in various benchmarks and relevance to current research trends. We aim to provide a robust and comprehensive evaluation of our proposed method using these models.

4.2 Evaluation Metrics

We treated our multi-label multi-class classification task as multiple independent binary classification tasks. Therefore, we employed classic metrics

for binary classifiers such as precision, recall, and the F1 score. To comprehensively evaluate performance across all labels, we calculated the average of these metrics over all labels. The macro-average calculates the arithmetic mean of the metrics across all categories, treating each category equally regardless of its frequency. It provides the model's overall performance by evaluating its ability to discriminate across all categories independently. In contrast, the micro-average gives more weight to categories with a larger number of samples, effectively reflecting the model's performance relative to the actual distribution of categories. Both metrics are commonly employed to assess the model performance in multi-label classification tasks, and neither can be considered universally superior. Consequently, we compared and reported both macro-average and micro-average.

4.3 Prompt-Based Experiments

We conducted experiments to evaluate the effectiveness of prompt engineering using the Llama3-8B-Instruct model. Specifically, we tested five different prompts on the OPP-115 corpus, using a configuration with temperature of 0.6, top-p of 0.9, and top-k of 50. The experimental results are summarized in the Table 1, which includes F1 scores as well as macro- and micro-average scores.

The overall performance of the model was relatively poor. The poor performance for some categories, like 'Introductory/Generic' and 'Practice Not Covered', can be attributed to them being less clearly defined, making it challenging to assess based on individual sentences.[1] Also, the model's strong hallucination led to an excessive number of false positives, resulting in high recall but low precision rates and consequently poor F1 scores, especially for categories like 'Data Retention', 'Data Security', and 'Do Not Track'. For instance, a segment that mentions protecting user privacy, like "We are committed to protecting and respecting your privacy" was mistakenly classified as 'Data Security'. While it refers to the commitment to privacy protection, it does not describe specific security measures, therefore, it is not related to this concept. On the other hand, the model performed well on categories such as 'First Party Collection/Use', 'Third Party Collection/Use', and 'International/Specific Audiences', probably due to their being easier concepts.

Differences in performance were observed between the five prompts. Prompt 1, which only contains a task description, performed the worst, which is not surprising given it providing the least information. Prompt 2 adds explanations of concept categories, so the model can understand the concepts better. Prompts 3 and 4 show a significant improvement in a few-shot setting. Compared to Prompt 3, where one example per concept category is used, Prompt 4 includes two examples. However, the increase in the number of examples did not improve the results, which was unexpected, indicating more is not always better. Prompt 5 used the CoT approach, but it performed the second worst, which was also unexpected.

4.4 Temperature Experiments

Several factors, such as sampling methods, temperature, top-p and top-k, can significantly impact model performance. We conducted experiments to show the role of temperature. We employed Prompt 3 for these experiments because it achieved the best performance among all five prompts as reported in the previous subsection. Utilizing the Llama3-8B-Instruct model again, experiments were conducted on the OPP-115 corpus with top-p fixed at 0.9 and top-k set to 50, while the temperature was varied across 0.3, 0.6, and 0.9, including a greedy generation for comparison. The performance under different generation configurations are shown in Table 2. It indicates that the Llama3-8B-Instruct model exhibits limited sensitivity to temperature variations for the task of concern.

4.5 Fine-Tuning: Baseline Experiments

We conducted experiments to evaluate the capability of fine-tuning the smaller versions of the three selected mainstream open-source LLMs, Llama3-8B, Qwen1.5-7B, and ChatGLM3-6B, utilizing four privacy policy corpora, OPP-115, GoPPC-150, CAPP-130, and APPCP-100. We perform LoRA fine-tuning on an RTX 4090 machine, primarily because LoRA significantly reduces computational costs compared to full fine-tuning and it can achieve a performance comparable to full fine-tuning in many scenarios, and it usually performs better than other alternative fine-tuning methods (Hu et al., 2023).

For OPP-115, we selected the results of

[1] As reported in (Wilson et al., 2016), during the annotation process, they have shown significant disagreement among the three legal experts. The 'OTHER' class (which covers the two aforementioned concepts) has the poorest inter-rater agreement, with Fleiss' kappa equal to just 0.49.

Label	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5
First Party Collection/Use	0.740	0.774	0.762	0.788	0.748
Third Party Collection/Use	0.730	0.772	0.762	0.714	0.758
User Choice/Control	0.366	0.441	0.465	0.458	0.478
User Access, Edit and Deletion	0.533	0.582	0.667	0.646	0.611
Data Retention	0.135	0.217	0.385	0.300	0.204
Data Security	0.549	0.517	0.549	0.550	0.471
Policy Change	0.472	0.467	0.512	0.568	0.532
Do Not Track	0.240	0.300	0.286	0.222	0.240
International/Specific Audiences	0.451	0.768	0.803	0.835	0.762
Introductory/Generic	0.436	0.564	0.431	0.471	0.514
Privacy Contact Information	0.731	0.696	0.707	0.682	0.714
Practice Not Covered	0.091	0.198	0.250	0.220	0.193
Macro Average	0.456	0.525	**0.548**	0.538	0.519
Micro Average	0.548	0.620	**0.636**	0.623	0.611

Table 1: Performance of Llama3-8B-Instruct using 5 prompts on the OPP-115 corpus (F1 scores).

Setting	Macro Average	Micro Average
Greedy	0.536	0.629
T=0.3	0.546	0.632
T=0.6	**0.548**	**0.636**
T=0.9	0.541	0.634

Table 2: Effect of temperature on the model performance (F1 scores).

PrivBERT (Srinath et al., 2021) as the baseline. For GoPPC-150, we adopted the PrivBERT+NN (neural network) approach used in (Tang et al., 2024) as the baseline. For CAPP-130, we used RoBERTa as the baseline because it achieved best performance among all models as described in (Zhu et al., 2023). For APPCP-100, we employed BERT+RF (random forests) described in (Zhang et al., 2024) as the baseline.

Table 3 presents the performances of different LLMs compared with the baselines, demonstrating a significant improvement[2]. Notably, the advantages of LLMs are more pronounced on GoPPC-150, suggesting their superiority in handling such complex and fine-grained tasks. Llama3-8B demonstrates a superior performance on the two English corpora but a slightly lower performance on the two Chinese corpora, likely due to the limited coverage of Chinese in its pre-training corpus. Due to the importance of GPT series in the field of LLMs, we also experimented with it. We fine-tuned and evaluated gpt3.5-turbo-0125 on the OPP-115 corpus. The final performance, with a macro-average F1 score of 0.801 and a micro-average F1 score of 0.851, shows no improvement over other three open-source models. Due to the prohibitive costs of fine-tuning GPT and its lack of significant performance advantages compared to other LLMs, we strategically limited our experimental evaluation of GPT to a single corpus (OPP-115).

4.6 Fine-Tuning: Experiments on Different Model Sizes

Prior research (Wei et al., 2022b) has indicated that, for some tasks, larger LLMs exhibit a significantly superior performance compared to smaller ones. However, in certain tasks, smaller models have been observed to achieve a substantial portion of the performance of larger models, thereby offering a more cost-effective and practical alternative. We conducted experiments to investigate this phenomenon for the task we are studying. We focused on the Qwen1.5 series, which encompasses a more diverse range of model scales especially at the lower end, including 0.5B, 1.8B, 4B, and 7B parameters. Qwen1.5-7B has also demonstrated a robust performance in our own experiments and other researchers' past studies, making it a reasonable choice.

We evaluated the performance of Qwen1.5 models of various scales (0.5B, 1.8B, 4B, and 7B) on all the four privacy policy corpora. Figure 3 shows the performances of Qwen1.5 models with different size, revealing a trend that larger models generally outperform smaller ones, but the improvement is small or marginal. Notably, after fine-tuning, the 0.5B model was already able to achieve over 90% of the performance of the 7B model. Therefore, if

[2]We also conducted experiments using Llama3.1 and Qwen2.5 on OPP-115 corpus. But it shows no improvement compared to Llama3 and Qwen1.5. So we did not employ them on other corpora. Llama3.1-8B: macro F1 0.828, micro F1 0.871; Qwen2.5-7B: macro F1 0.825, micro F1 0.872.

Standard		Baseline		Llama3-8B		Qwen1.5-7B		Chatglm3-6B	
		macro	micro	macro	micro	macro	micro	macro	micro
OPP-115	All	0.830	0.870	**0.836**	**0.877**	0.831	0.868	0.819	0.867
GoPPC-150	Level 1	0.669	0.697	**0.717**	**0.725**	0.709	0.718	0.705	0.712
	All	0.529	0.589	**0.618**	**0.685**	0.612	0.673	0.609	0.668
CAPP-130	Topic	0.841	0.819	0.838	0.821	0.852	0.829	**0.858**	**0.837**
APPCP-100	Level 1	0.832	0.867	0.831	0.875	**0.843**	**0.883**	0.840	0.878
	All	0.767	0.846	0.767	0.858	**0.782**	**0.865**	0.778	0.858

Table 3: Performance of different LLMs with fine-tuning compared with baseline.

the performance requirements are not high, smaller models have certain advantages.

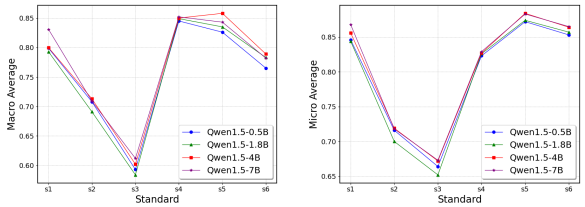

Figure 3: Effect of model size on the performance. Standards s1-s6 represent OPP-115, GoPPC-150 level-1, All nodes, CAPP-130 Topic, APPCP-100 level-1 and All nodes, respectively.

4.7 Fine-Tuning: Experiments on Single- vs Multi-Task Settings

Single-task training focuses on optimizing a model for one specific task, while multi-task training involves training a model on multiple tasks simultaneously. We conducted experiments to explore the application of multi-task fine-tuning to enhance the ease of model deployment. Specifically, we integrated the first- and second-level node classification tasks of GoPPC-150 into a unified task framework by merging the two aforementioned single-task fine-tuning corpora. By fine-tuning the Llama3-8B model on this consolidated corpus, we achieved an excellent performance on both tasks simultaneously. Table 4 presents the performances of the two training paradigms. The performance of multi-task fine-tuning shows just a small significant drop compared to single-task fine-tuning, demonstrating its feasibility.

4.8 Performance Comparison

As mentioned in Section 2, Goknil et al. (2024) explored the use of prompt engineering and LLMs for automated privacy policy analysis, using OPP-115 as the corpus. To compare with their work, we adopted their best results on Llama3-8B. Since they

Method	Macro average	Micro average
Single-task	0.618	0.685
Multi-task	0.614	0.679

Table 4: Performances of single- and multi-task paradigms.

did not consider the last three categories in OPP-115, we also excluded them in our performance comparison experiments, leading to slight discrepancies with the results reported in Section 4.3. Table 5 presents a comparison of the performance figures (F1 scores) on the OPP-115 corpus. The results show that prompt engineering methods generally perform more poorly, while the fine-tuning method we employed demonstrates a far superior performance.

Category	Goknil et al.'s	Ours (PE)	Ours (Finetuned)
First Party Collection/Use	0.760	0.789	**0.939**
Third Party Sharing/Collection	0.710	0.789	**0.935**
User Choice/Control	0.630	0.477	**0.847**
User Access, Edit and Deletion	0.730	0.667	**0.821**
Data Retention	0.400	0.348	**0.696**
Data Security	0.740	0.568	**0.873**
Policy Change	0.880	0.585	**0.973**
Do Not Track	0.810	0.300	**1.000**
International and Specific Audiences	0.810	0.827	**0.918**
Micro Average	0.730	0.694	**0.916**

Table 5: Comparison of performances (F1 scores) on the OPP-115 corpus (PE = prompt-engineered).

5 Explainability

Compared to traditional deep learning methods, LLMs has the potential to offer enhanced explainability due to its capability of producing human-

like texts in natural languages. To investigate the explainability of LLMs for privacy policy concept classification, we fed privacy policy segments along with their corresponding concept categories into an LLM, prompting it to analyze and explain the classification results. Specifically, our prompts include the task description, the concept categories' descriptions, the required output format, and some examples. The task description and concept categories' descriptions are consistent with those detailed in Section 3.2. We instructed the LLM to, for each category, first explain its meaning and then analyze the segment's relevance to the category.

As an example, we utilized the Llama3-8B-Instruct model, with settings of temperature=0.6 and top-p=0.9, to generate explanations for 100 privacy policy segments randomly selected from OPP-115. We focused on the first 11 categories of the OPP-115 taxonomy, excluding the 'Practice Not Covered' category, which does not require any specific explanation.

We established three metrics to assess the quality of the LLM-generated explanations, as explained below. Each metric was scored by three human annotators on a scale of 1, 2, or 3, where 1 indicates 'poor performance', 2 indicates 'acceptable performance', and 3 indicates 'outstanding performance'.

Completeness assesses whether the explanation covers all key points of the privacy policy segment that identifies the relevant concept categories.

Logicality evaluates the accuracy of the model's understanding of the privacy policy segment and the coherence of the model's reasoning.

Comprehensibility focuses on the clarity and understanding of the explanation itself, especially in terms of language.

The three human annotators are three co-authors of the paper, all postgraduate research students, who conducted a qualitative evaluation of the explanations of LLM outputs based on the three metrics mentioned above. To prevent potential positive scoring bias, we included 10 decoy explanations that were made blind to the annotators. These decoy explanations were crafted to exhibit at least one aspect of relatively poor performance while maintaining basic explanatory quality. After aggregating the scores, the average scores for the three metrics are presented in Table 6. The average scores of the 10 artificially crafted explanations are significantly lower across all three metrics compared to the LLM-generated ones. We assessed the inter-rater reliability among three annotators using Fleiss' kappa (Landis and Koch, 1977). The results indicated a substantial agreement for all three metrics: 0.765 for completeness, 0.695 for logicality, and 0.656 for comprehensibility. Our primary finding is that LLMs exhibit very very good explainability in explaining the classification results of the 100 privacy policy segments, across all three metrics. Notably, the Llama3-8B-Instruct model tend to offer comprehensive analyses of the original text, which contributes to their strong performance in terms of completeness. Moreover, the language style of the LLM-generated content can be easily set to be clear, concise, and easy to understand through the use of prompts, thus demonstrating strong comprehensibility. However, the Llama3-8B-Instruct model' understanding of privacy policy segments occasionally lacks depth, which results in slightly lower logicality score.

Source	C1	L	C2
LLM	2.84	2.73	2.87
Artificial	2.43	2.10	2.73

Table 6: Average scores of LLM-generated explanations and artificially crafted explanations (C1 = completeness, L = Logicality, C2 = comprehensibility).

6 Conclusion

This paper proposes a method for utilizing LLMs to classify concepts in a privacy policy based on an established taxonomy. Unlike prior studies, we provided a comprehensive evaluation of LLMs in this domain, incorporating both prompt engineering and LoRA techniques, and assess performance across four SOTA privacy policy corpora and multiple mainstream LLMs, achieving SOTA results against existing methods. We investigated the effects of factors such as temperature, model size, and training paradigm. To enhance the explainability of the classification results, we used LLMs to generate explanations for the identified concepts and designed an evaluation framework to assess the explanations based on three metrics: completeness, logicality, and comprehensibility. The findings demonstrate that LLMs can provide satisfactory explanations to three human annotators. This paper highlights the great potential of LLMs for both automating analysis of privacy policies and producing useful human-understandable explanations, therefore opening up their use for many downstream tasks in this important application domain.

7 Limitations

The performance achieved using prompt engineering in our experiments is quite poor. This can be two reasons: more advanced prompt engineering methods are necessary, and LLMs may not have seen enough privacy policies so fine-tuning is a must to improve the performance of any tasks about privacy policies. We call for more follow-up research to clarify both points.

Due to resource limitation, we primarily utilized smaller and locally deployed models. While these models achieved SOTA performances in our experiments and demonstrated a great potential, we did not experiment with larger models like Llama-3-70B and GPT-4. As a result, we were unable to evaluate the performance upper bound of LLMs for privacy policy concept classification. Additionally, we achieved promising results using LoRA. Prior studies (Hu et al., 2022) have shown that LoRA can closely approximate the performance of full-parameter fine-tuning. However, the underlying mechanisms of these approaches differ, and further research is required to fully explore the potential of full-parameter fine-tuning.

LLMs also benefit from pre-training to acquire domain-specific knowledge. Some studies (Gupta et al., 2023; Ke et al., 2023) adopted the continual pre-training paradigm, enabling models to perform unsupervised learning on domain-specific corpora before being fine-tuned for specific tasks. This approach allows LLMs to acquire substantial knowledge in a given domain and therefore likely to be able to solve targeted problems more effectively. In this paper, we did not adopt the continual pre-training paradigm, but relied on fine-tuning to help LLMs learn domain-specific knowledge. The effectiveness of the continual pre-training paradigm remains an area for future research.

In order to support other researchers to reproduce our results and to conduct follow-up research, we will make all data and code used publicly available.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, Florian Schaub, and Norman Sadeh. 2020. Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text. In *Proceedings of The Web Conference 2020*, pages 1943–1954. ACM.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

California State Legislature, USA. 2018. California Consumer Privacy Act of 2018. Cal. Civ. Code §§ 1798.100-1798.199.

Orlando Amaral Cejas, Sallam Abualhaija, and Lionel C. Briand. 2024. CompAi: A tool for GDPR completeness checking of privacy policies using artificial intelligence. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pages 2366–2369. ACM.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):39:1–39:45.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang,

Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. ChatGLM: A family of large language models from glm-130b to GLM-4 all tools. *Preprint*, arXiv:2406.12793.

Arda Goknil, Femke B. Gelderblom, Simeon Tverdal, Shukun Tokas, and Hui Song. 2024. Privacy policy analysis through prompt engineering for LLMs. *Preprint*, arXiv:2409.14879.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pretraining of large language models: How to re-warm your model? In *Proceedings of the 2023 Workshop on Efficient Systems for Foundation Models at the 40th International Conference on Machine Learning*.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *Proceedings of the 27th USENIX Security Symposium*, pages 531–548. USENIX Association.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 2022 International Conference on Learning Representations*, page 13.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. 2023. LLM-Adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5276. ACL.

Duha Ibdah, Nada Lachtar, Satya Meenakshi Raparthi, and Anys Bacha. 2021. "Why should I read the privacy policy, I just need the service": A study on attitudes and perceptions toward privacy policies. *IEEE Access*, 9:166465–166487.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pretraining of language models. In *Proceedings of the 11th International Conference on Learning Representations*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. ACL.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597. ACL.

Najmeh Mousavi Nejad, Pablo Jabat, Rostislav Nedelchev, Simon Scerri, and Damien Graux. 2020. Establishing a strong baseline for privacy policy classification. In *ICT Systems Security and Privacy Protection: 35th IFIP TC 11 International Conference, SEC 2020, Maribor, Slovenia, September 21–23, 2020, Proceedings*, pages 370–383. Springer.

Majd Mustapha, Katsiaryna Krasnashchok, Anas Al Bassit, and Sabri Skhiri. 2020. Privacy policy classification with XLNet (short paper). In *Data Privacy Management, Cryptocurrencies and Blockchain Technology: ESORICS 2020 International Workshops, DPM 2020 and CBT 2020, Guildford, UK, September 17–18, 2020, Revised Selected Papers*, volume 12484 of *Lecture Notes in Computer Science*, pages 250–257. Springer.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2024. The prompt report: A systematic survey of prompting techniques. *Preprint*, arXiv:2406.06608.

Mukund Srinath, Shomir Wilson, and C. Lee Giles. 2021. Privacy at scale: Introducing the PrivaSeer corpus of web privacy policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6829–6839. ACL.

Peng Tang, Xin Li, Yuxin Chen, Weidong Qiu, Haochen Mei, Allison Holmes, Fenghua Li, and Shujun Li. 2024. A comprehensive study on GDPR-oriented analysis of privacy policies: Taxonomy, corpus and GDPR concept classifiers. *Preprint*, arXiv:2410.04754.

Damiano Torre, Sallam Abualhaija, Mehrdad Sabetzadeh, Lionel Briand, Katrien Baetens, Peter Goes, and Sylvie Forastier. 2020. An AI-assisted approach

for checking the completeness of privacy policies against GDPR. In *Proceedings of the 2020 IEEE 28th International Requirements Engineering Conference*, pages 136–146. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Paul Voigt and Axel Von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *Proceeings of the 2022 International Conference on Learning Representations*, page 46.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022c. Chain of thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340. ACL.

Anhao Xiang, Weiping Pei, and Chuan Yue. 2023. PolicyChecker: Analyzing the GDPR completeness of mobile apps' privacy policies. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3373–3387. ACM.

Xiheng Zhang, Xin Li, Peng Tang, Ruiqi Huang, Yuan He, and Weidong Qiu. 2024. Privacy policy compliance detection and analysis based on knowledge graph. GitHub repository.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):20:1–20:38.

Pengyun Zhu, Long Wen, Jinfei Liu, Feng Xue, Jian Lou, Zhibo Wang, and Kui Ren. 2023. CAPP-130: A corpus of chinese application privacy policy summarization and interpretation. In *Advances in Neural Information Processing Systems*, volume 36, pages 46773–46785. Curran Associates Inc.

A Details of Five Types of Prompts

As mentioned in Section 3.2, we designed five types of prompts, and Figure 4 shows an example of these five prompts.

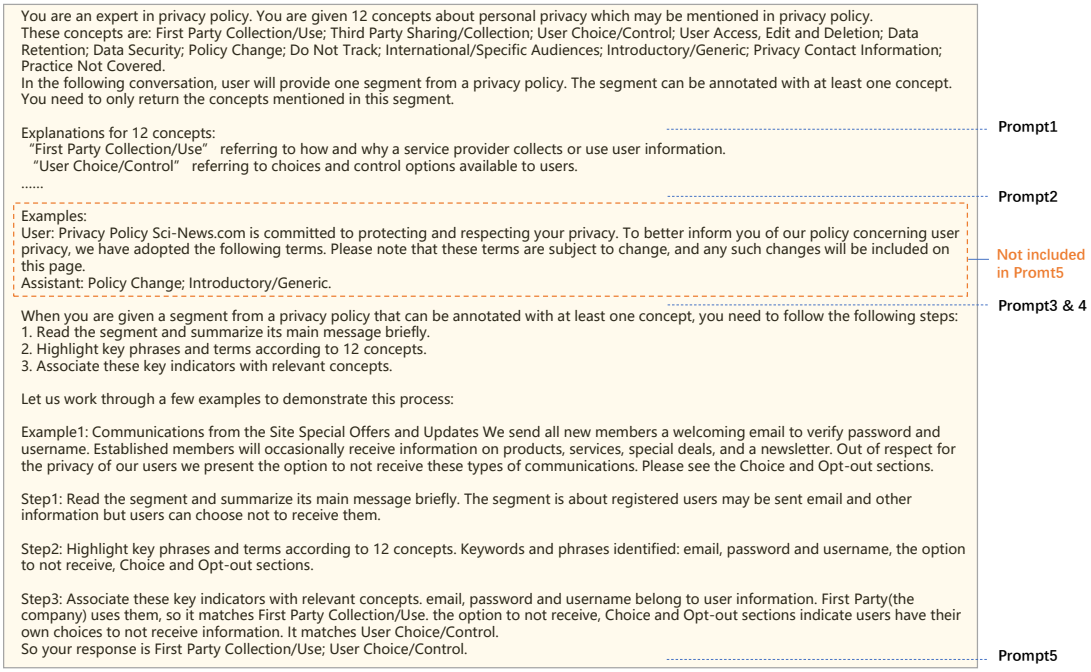

Figure 4: Examples of the five types of prompts used in our experiments. The content within the yellow dashed box is not included in Prompt 5.