Privacy Checklist: Privacy Violation Detection Grounding on Contextual Integrity Theory

Haoran Li[1]*, **Wei Fan**[1]*, **Yulin Chen**[2], **Jiayang Cheng**[1]
Tianshu Chu[3], **Xuebing Zhou**[3], **Peizhao Hu**[3], **Yangqiu Song**[1]
[1]HKUST, [2]National University of Singapore, [3]Huawei Technologies
{hlibt, wfanag, jchengaj}@connect.ust.hk, chenyulin28@u.nus.edu
{chutianshu3, Xuebing.Zhou, hu.peizhao}@huawei.com, yqsong@cse.ust.hk

Abstract

Privacy research has attracted wide attention as individuals worry that their private data can be easily leaked during interactions with smart devices, social platforms, and AI applications. Existing works mostly consider privacy attacks and defenses on various sub-fields. Within each field, various privacy attacks and defenses are studied to address patterns of personally identifiable information (PII). In this paper, we argue that privacy is not solely about PII patterns. We ground on the Contextual Integrity (CI) theory which posits that people's perceptions of privacy are highly correlated with the corresponding social context. Based on such an assumption, we formulate privacy as a reasoning problem rather than naive PII matching. We develop the first comprehensive checklist that covers social identities, private attributes, and existing privacy regulations. Unlike prior works on CI that either cover limited expert annotated norms or model incomplete social context, our proposed privacy checklist uses the whole Health Insurance Portability and Accountability Act of 1996 (HIPAA) as an example, to show that we can resort to large language models (LLMs) to completely cover the HIPAA's regulations. Additionally, our checklist also gathers expert annotations across multiple ontologies to determine private information including but not limited to PII. We use our preliminary results on the HIPAA to shed light on future context-centric privacy research to cover more privacy regulations, social norms and standards.

1 Introduction

As machine learning models and their applications achieve wide accessibility and applicability, they revolutionize our daily lives. On the one hand, these data-driven models achieve consistent utility improvements with increasing training corpus. On the other hand, private data may be collected unintentionally, which leads to individuals' increasing privacy concerns. To achieve trustworthy AI, great efforts have been made to study privacy issues with regard to emerging applications (Li et al., 2023a; Debenedetti et al., 2023).

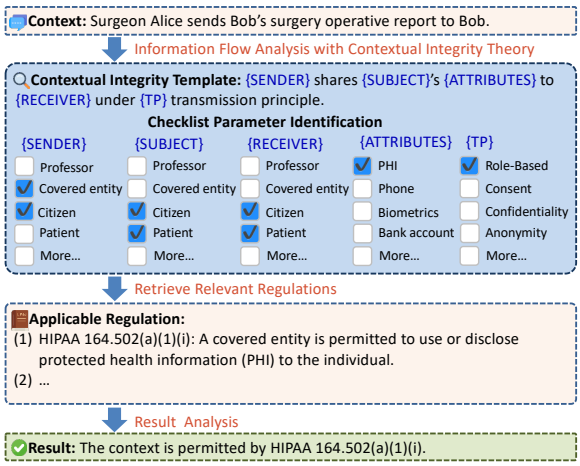

Figure 1: The exemplary case to transform privacy issues into reasoning problems. Our proposed checklist collects roles, attributes, transmission principles and annotated legal norms to facilitate the reasoning process.

Existing privacy studies conducted by researchers are relatively fragmented. Each field has its own privacy formulation based on specific tasks. For example, in computer networks, network communications' confidentiality is widely discussed (Nath Nayak and Ghosh Samaddar, 2010; Fonseca et al., 2009) and cryptographic methods are proposed to protect cybersecurity (Buehrer et al., 2005; Needham and Schroeder, 1978). When it comes to deep learning, in both Computer Vision and Natural Language Processing fields, there is a significant body of research on privacy attacks (Shokri et al., 2016; Carlini et al., 2021; Li et al., 2023b) and defenses (Abadi et al., 2016).

In addition to the fragmented formulations, the current scope of privacy research is narrow. Most existing privacy studies focus on protecting private attributes specified by privacy regulations. Though these sensitive attributes are indeed crucial components of privacy regulations, they only constitute a

*Haoran Li and Wei Fan contributed equally.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1748–1766
April 29 - May 4, 2025

small fraction of regulations' coverage. However, privacy is not solely about protecting attributes. Instead, It involves understanding the context to ensure compliance with existing regulations and norms. For example, according to the Health Insurance Portability and Accountability Act of 1996 (HIPAA), protected health information (PHI) is considered sensitive and should be protected. However, as shown in Figure 1, though Bob's surgery operative report is classified as protected health information, if we identify Alice as a covered entity and Bob as the individual (patient), we may apply HIPAA 164.502(a)(1)(i) to permit surgeon Alice to share Bob's PHI with him. This example demonstrates that privacy issues can be resolved through reasoning within the context. When we consider privacy, we actually evaluate whether our information transmission contexts violate existing standards and expectations.

To bridge the gap, Contextual Integrity (CI) theory (Nissenbaum, 2010) is proposed to examine the appropriateness of privacy information flows based on specific contexts. A few prior works on CI resort to formal languages such as first-order logic (Barth et al., 2006; Benthall et al., 2017; Shvartzshnaider et al., 2016) to fit contexts into pre-defined clauses to determine privacy violations. However, it requires heavy effort for experts to annotate complicated legal documents into logical language. Consequently, none of these approaches can scale up to cover even a whole document.

In this paper, to leverage the reasoning capability of privacy information flows while allowing substantial flexibility to handle natural language ambiguity and variability, we present *Privacy Checklist*. We formulate the privacy evaluation as an in-context reasoning problem with existing large language models (LLMs). Instead of annotating separate clauses, our *Privacy Checklist* uses large language models to extract CI characteristics and use the tree structure to capture their hierarchical nature. In addition, our *Privacy Checklist* also includes role-based and attribute-based graphs, as well as a definition dictionary to facilitate in-context reasoning. To determine privacy violations for the given context, we first fetch relevant regulations and use large language models to conduct in-context reasoning with retrieved regulation. We annotate all information transmission norms inside the HIPAA and show that our *Privacy Checklist* can improve 6% - 18% accuracy on average for existing LLMs to improve their privacy judgment ability in real court cases. Our contribution can be summarized as follows[1]:

1) We extend prior works on CI to natural language and formulate the privacy issue as an in-context reasoning problem with the help of LLMs.

2) We present *Privacy Checklist*, a first scalable knowledge base that can cover all norms of information transmission inside the HIPAA.

3) We consider various retrieval-augmented generation (RAG) pipelines. To retrieve relevant legal documents, we implement term frequency, semantic similarity, and agent-based methodologies.

4) We conduct comprehensive experiments to show that our checklist is effective in improving LLMs' privacy judgment for court cases.

2 Preliminaries

Contextual Integrity, norms and applications. Contextual Integrity (Nissenbaum, 2010) posits that privacy is related to information flows and information flows must adhere to the established norms of the context to protect privacy. Each norm can typically be viewed as an atomic unit of information transmission involving three key actors: the sender, the receiver, and the data subject. Depending on the social context of these actors, such as their roles and the disclosed attributes of the data subject, we can determine whether the norm is positive or negative, which means the norm is either permitted or prohibited by existing privacy regulations. Previous works on CI (Barth et al., 2006; Benthall et al., 2017; Shvartzshnaider et al., 2016) aimed to transform the context into explicit formal languages, such as first-order logic, to determine privacy violations with rule-based frameworks. However, as shown in Figure 2 (a), it costs heavily for experts to annotate such a knowledge base of norms. Experts must meticulously define a set of predicates and annotate the entire body of regulations, subpart by subpart. Consequently, these works on formal languages fail to scale up their norms due to the inherent complexity of legal documents. On the other hand, existing legal LLMs, including LawGPT (Zhou et al., 2024), Lawyer LLaMA (Huang et al., 2023) and ChatLaw (Cui et al., 2023) are proposed on the general legal domain and underperform on the legal domain due to data scarcity. GoldCoin (Fan et al., 2024) proposed instruction tuning grounded on CI to en-

[1]Code is publicly available at https://github.com/HKUST-KnowComp/privacy_checklist.

hance LLMs' judgment ability. Instead, our work reveals the flaws of synthetic data and shows that the proper RAG pipelines without further tuning can yield comparable or even better performance.

In addition, for discussions about current privacy attacks and defenses, please refer to Appendix C.

3 Privacy Checklist Construction

In this section, we show how to construct our *Privacy Checklist* based on existing privacy regulations with less annotation cost. For the given regulation, our *Privacy Checklist* encompasses a document tree with CI characteristics, a role graph, an attribute graph, and a definition dictionary.

3.1 Legal Document Processing

We start by crawling the full HIPAA document from the official Code of Federal Regulations website. We implement a parser with regular expressions to capture the regulation identifiers and corresponding contents. Since the legal documents are rigorously written and well-organized, we can represent the HIPAA rules' subsumption relationships using a tree structure $\mathcal{T} = \{\mathcal{V}, \mathcal{E}\}$ where edges in $\mathcal{E}$ indicates the subsumption relationship between two nodes. For nodes in $\mathcal{V}$, leaf nodes correspond to specific, non-separable specifications (e.g., 164.502(a)(1)(i)), while internal nodes denote broader subparts (e.g., 164.502(a)(1)). We set "HIPAA" as the dummy root node. In addition, we observe frequent cross-references among the nodes. For each node, any reference of other identifiers is recorded as the node attribute.

3.2 CI Characteristics Annotation

For each leaf, we concatenate all texts from the root to the current leaf node to obtain its full specification. Then, we exploit GPT-4 as the annotator to execute the following tasks:

Specification classification. Due to the complexity of legal documents, each leaf node's specification can either be a regulatory norm or a clarification of the norm. Such clarifications are irrelevant to information transmission but play crucial roles in regulatory norms. Therefore, for each specification, we ask GPT-4 to conduct a three-way classification to determine if the specification is a positive norm, negative norm, or general definition. A positive norm denotes the information transmission action (norm) that is permitted by law, while a negative norm prohibits such information transmission. Otherwise, the specification is regarded as the general definition, which enhances the norms' clarity.

CI characteristics extraction. After identifying all the positive and negative norms, we continue to extract characteristics identified by the Contextual Integrity Theory from each norm. First, we request GPT-4 to identify three actors' roles, including the sender role, receiver role, and subject role for the given norm. Then, we inquire GPT-4 to determine the subject's attributes that are disclosed during information transmission. Additionally, we ask GPT-4 to analyze additional context, including the purpose and consent form of transmission. Finally, we let GPT-4 specify whether the sender and the subject are the same person and, similarly, whether the receiver and the subject are the same person.

Reference annotation. Previous works, as shown in Figure 2 (a), commonly treat each norm as an isolated individual and discard the cross-references among norms. However, legal regulations are rather complicated, and referring to these references is essential to assessing privacy violations accurately. For instance, HIPAA 164.502(a)(1)(iv) states that "A covered entity is permitted to use or disclose protected health information except for uses and disclosures prohibited under 164.502(a)(5)(i), pursuant to and in compliance with a valid authorization under 164.508." To permit data transmission with 164.502(a)(1)(iv), we must first ensure that no restrictions apply under 164.502(a)(5)(i) and then confirm that the authorization (one type of consent form) is valid under 164.508. Therefore, we also extend our checklist to include the annotated references for all norms in $\mathcal{T}$. For each norm, we prompt its content to GPT-4 and ask about the relationship between references inside the node attribute and the specified norm. Each reference is then categorized as either a support or an exception of the given norm.

3.3 Auxiliary Knowledge Collection

In addition to norm annotation, most regulations also bring their own terminologies and use these terms throughout their content. To facilitate the reasoning process and align the given context with corresponding regulations, we further collect two auxiliary graphs and one definition dictionary.

Role graph $\mathcal{G}^r$. Role instances extracted from daily context frequently misalign with the abstract

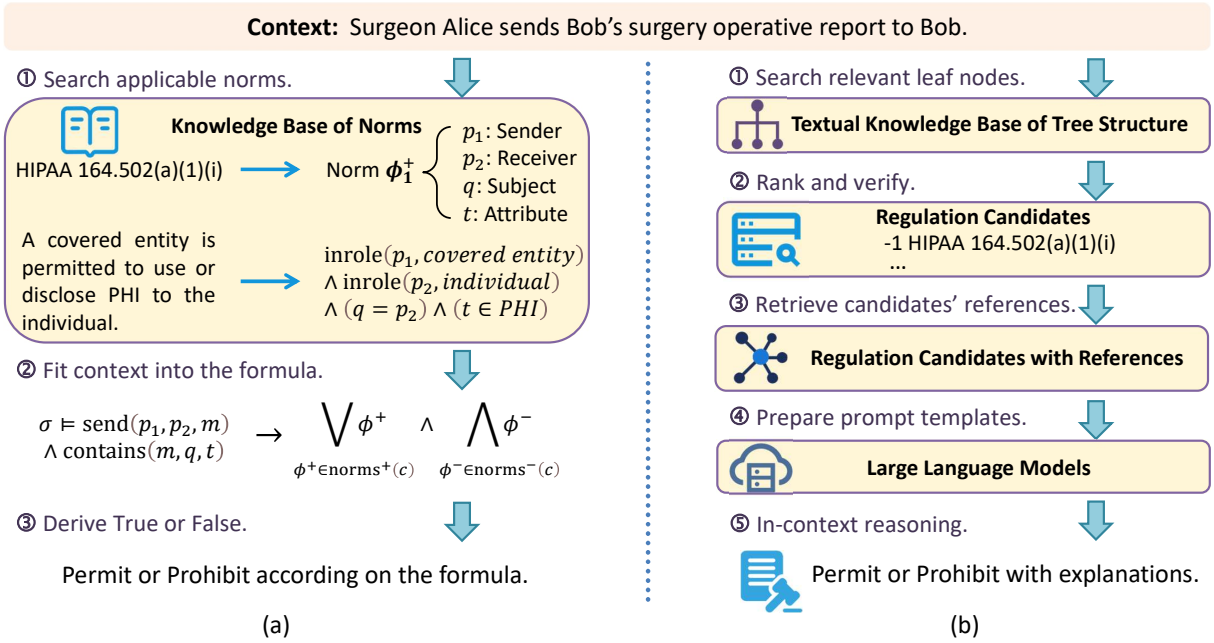

Figure 2: The overview of privacy reasoning within the given contexts. Subfigure (a) illustrates previous approaches that use formal languages to determine privacy violations based on rules of inference and axioms. Instead, in subfigure (b), we propose an in-context reasoning pipeline with our proposed *Privacy Checklist* and LLMs.

roles defined in regulations. For example, in HIPAA, the term covered entity may refer to 1) a health plan, 2) a health care clearinghouse or 3) a health care provider. To permit Alice's action in Figure 1, we need to link Alice's role, surgeon, to the covered entity. Based on this observation, we further curate a role graph $\mathcal{G}^r = \{\mathcal{V}^r, \mathcal{E}^r, \mathcal{R}^r\}$ where vertices $v \in \mathcal{V}^r$ represent social roles and only two relations "subsume" and "is subsumed by" are in $\mathcal{R}^r$. To construct our role graph $\mathcal{G}^r$, we align the roles in the legal documents with the WordNet (Miller, 1994). We first use GPT-4 and parsers to recognize social roles in the legal document, then we align these roles in the WordNet and recursively append their hypernyms until the role root "person.n.01" is reached. For new definitions of roles such as the covered entity, if these definitions include roles in $\mathcal{V}^r$, we further append these new definitions as roles.

Attribute graph $\mathcal{G}^a$. Similarly, to identify whether the transmitted attributes are sensitive, we need an attribute graph $\mathcal{G}^a = \{\mathcal{V}^a, \mathcal{E}^a, \mathcal{R}^a\}$ where vertices $v \in \mathcal{V}^a$ represent private attributes and only two relations "subsume" and "is subsumed by" are in $\mathcal{R}^r$. With $\mathcal{G}^a$, we can map the surgery operative report to the protected health information (PHI) specified by HIPAA. To collect private attributes protected by existing regulations, we incorporate expert annotated ontologies, including Data Privacy Vocabulary (Pandit et al., 2019) and OPPO (Gupta and Hahmann, 2023) into $\mathcal{G}^a$. Our $\mathcal{G}^a$ gathers the ontologies' inclusion relationships for both class-class and class-individual. Then, for new terms that include attributes in $\mathcal{V}^a$, we further append them into $\mathcal{G}^a$.

Definition dictionary $\mathcal{D}$. To further facilitate in-context reasoning, we also implement an HTML parser to build a definition dictionary $\mathcal{D}$ based on legal documents' definition chapters. Keys of $\mathcal{D}$ are introduced terms and terms' corresponding definitions are values.

3.4 Data Summary and Manual Validation

Data statistics. In our collected *Privacy Checklist*, our processed document tree $\mathcal{T}$ includes 642 internal nodes, 1,673 leaf nodes, 2,098 edges for subsumption relationship and 761 cross references. For the annotated privacy rules, we identify 231 positive norms (permitted information transmission norms), 32 negative norms as well as 328 general definitions of specifications for applying norms within $\mathcal{T}$. For our role graph $\mathcal{G}^r$, we collect 296 roles with 318 subsumption relations. For our attribute graph $\mathcal{G}^a$, we collect 211 sensitive attributes with 485 subsumption relations. Moreover, our definition dictionary $\mathcal{D}$ encompasses 54 definitions formally defined in the HIPAA 164.103.

Expert verification. To ensure the quality and correctness of our *Privacy Checklist*, two authors and one law student manually validate positive and

negative norms with the original regulations and rectify the inconsistently parsed CI characteristics.

4 Enhance Reasoning with Checklist

In this section, we show how to exploit LLMs' in-context reasoning ability to determine privacy violations. As shown in Figure 2 (b), the basic intuition is to use existing LLMs for a retrieval-augmented generation (RAG) pipeline.

4.1 Context Annotation Methods

In addition to directly feeding the whole context for retrieval, we adopt the following annotation methods for a given context:

CI characteristics extraction. We follow the same CI characteristics extraction procedures used for building our checklist to extract the sender role, receiver role, subject role, transmitted attribute, purpose, and consent form from the given context.

LLM explanation. There are often discrepancies between the CI characteristics extracted from legal documents and those derived from real-world contexts. Typically, CI characteristics from legal documents are defined using exclusive legal terminology, whereas those from daily contexts tend to be more specific and practical. Without the knowledge about definitions in $\mathcal{D}$, existing retrieval methods struggle to accurately search for relevant rules based on the given context. For example, both statistical and semantic methods of information retrieval fail to link the given role query "surgeon" to "covered entity". To address this, we leverage LLMs to explain the context using terms inside designated laws. After the explanation, the retrieval process can be more accurate.

4.2 Rule Retrieval Methods

To retrieve relevant regulations from our *Privacy Checklist*, we implement the following retrieval methods based on the parsed context:

BM25. BM25 retrieves documents based on the term frequency-inverse document frequency (TF-IDF) score (Salton and Buckley, 1988) and word frequency. We use the LLM explanation as the query Q, aiming to select the most relevant regulations by measuring the similarity between the LLM explanation query Q and the regulation texts E in the document tree $\mathcal{T}$. For a word w in the query Q, the relationship between the word and the regulation definition is calculated as follows:

$$s(w,E) = \frac{\mathrm{IDF}(w) \cdot f(w,E) \cdot (k_1+1)}{f(w,E) + k_1 \cdot (1 - b + b \cdot \frac{|E|}{\mathrm{avgdl}})}, \quad (1)$$

where $\mathrm{IDF}(w)$ represents the inverse document frequency of the word w, $f(w,E)$ denotes the frequency of the word w appearing in E, $|E|$ is the length of E, and avgdl is the average length across the subrules in $\mathcal{T}$. k_1 and b are hyperparameters, and we set $k_1 = 1.5$ and $b = 0.75$. The similarity between a query Q and the definition E is calculated by summing the scores of words in Q:

$$Sim(Q,E) = \sum\nolimits_{w_i \in Q} s(w_i, E). \quad (2)$$

The regulations with the highest scores are then selected as the retrieved results.

Embedding similarity. In addition to statistical similarity, we also implement retrieval methods based on semantic similarity. Specifically, we use pretrained Sentence Transformers (Reimers and Gurevych, 2019) to calculate both contexts and regulations' embeddings. Then, the cosine similarities are calculated for embeddings of the contexts and regulations. Notably, besides calculating semantic similarity using chunks of text, we can also calculate semantic similarity on roles and attributes. These roles and attributes can then be used as keywords to retrieve applicable rules. For simplicity, we use "all-mpnet-base-v2" (Song et al., 2020) as our embedding model throughout the paper.

Agent-based retrieval. Besides directly retrieving relevant rules from our *Privacy Checklist*, we consider LLMs themselves as knowledge bases (Petroni et al., 2019) that have preliminary legal knowledge. Based on this heuristic, we implement agent-based retrieval by treating powerful LLMs as lawyers and use prompt engineering to generate applicable rules with corresponding IDs. To mitigate misinformation and hallucination, the agent-based retrieval further leverages our *Privacy Checklist* to verify if the generated IDs are valid.

4.3 In-context Reasoning

With the retrieved rules, we may leverage existing LLMs to perform in-context reasoning. For each candidate rule, we apply a Chain-of-Thought (CoT) prompting template (Wei et al., 2022; Wang et al., 2023) to assess its applicability in judging privacy violations within a given context. Our CoT prompt template includes four components. First, the LLM

is instructed to identify the information transmission flow and analyze the three stakeholders, as well as the transmitted attributes and purpose from the given context. Second, the LLM reviews the analyzed information and determines if it matches the candidate rule's legal terms. Third, the LLM assesses whether the candidate rule is relevant to the context. Lastly, the LLM decides whether the context is in compliance, in violation, or unrelated to the candidate rule.

5 Experimental Setups

Data. We use collected real and synthetic court cases for evaluation to demonstrate the usefulness of our *Privacy Checklist*. Following Fan et al. (2024), we use their collected real and synthetic court cases about HIPAA. The real cases are collected from the Caselaw Access Project (CAP)[2] and the synthetic cases are LLM-augmented data based on real cases and HIPAA sub-rules. For both datasets, we directly prompt existing LLMs to perform three-way classification by determining if the given case is *permitted* by, *prohibited* by or *not applicable* to HIPAA. The dataset statistics are summarized in Table 1.

Evaluated LLMs. We evaluate both open-source and close-source LLMs. In terms of open-source LLMs, we download their model weights of instruction or chat versions and generate responses on two NVIDIA Tesla A100 40GB graphic cards. Our evaluated open-source models include Llama3 (AI@Meta, 2024), Qwen1.5, Qwen2 (Yang et al., 2024), ChatGLM4 (GLM et al., 2024), and Mistral-v0.3 (Jiang et al., 2023). For close-source LLM, we evaluate the GPT-4-turbo-04-09's performance with API accesses.

Evaluated methods. We evaluate LLMs' performance over the following methods. The first three methods rely on prompting tricks while the last three methods reason over our *Privacy Checklist* with different retrieval methods.

• Direct prompt (**DP**). We prompt LLMs with the case context and directly ask them to determine if the given context is permitted, prohibited, or unrelated to HIPAA.

• CoT prompt with automatic planning (**CoT-auto**). We prompt LLMs to automatically generate step-by-step planning to analyze the given case and then execute the steps to determine privacy violations.

• CoT Prompt with manual guidelines (**CoT-manual**). Instead of using LLMs to generate plans, we prompt LLMs with pre-defined guidelines for each step. Our guidelines follow the CI theory to instruct LLMs to analyze the CI characteristics step by step to assess privacy violations.

• CoT prompt with regulation IDs from *agent-based retrieval* (**Agent-ID**). We first ask LLMs with the case to generate applicable regulation IDs. Then, we verify their existence with our checklist to filter out misinformation. Finally, we prompt LLMs similarly to the *CoT-manual* approach, but with the addition of regulation IDs for in-context reasoning.

• CoT prompt with *LLM explanation* and regulations retrieved via BM25 (**BM25-content**). We apply the *LLM explanation* to clarify the case context with legal terms to facilitate the retrieval process. Then, we use BM25 to search for relevant sub-rules. Finally, we incorporate both content and IDs of these sub-rules into the *CoT-manual* prompt to improve in-context reasoning.

• CoT prompt with *CI characteristics extraction* and regulations retrieved via embedding similarity (**CI-ES-content**). We first perform *CI characteristics extraction* to extract necessary information about information flow. Then, we consider stakeholders' roles as keywords and use pre-trained embedding models to match their roles with roles in our *Privacy Checklist* by calculating the embedding similarities. We use the matched roles as keywords to search for applicable rules. Finally, we add the regulations to the *CoT-manual prompt* for in-context reasoning.

Evaluation metrics. For the three-way classification, we report the accuracy of correct LLM judgments in a single run. In addition, for each class, we also calculate its precision, recall, and F1. To verify the correctness, we implement multiple parsers based on regular expressions to capture the desired predictions. All parsing failures are regarded as incorrect predictions.

6 Experimental Results

In this section, we show how our *Privacy Checklist* enhances existing LLMs' privacy assessment ability via zero-shot in context reasoning.

[2] https://case.law/

Type	Permit	Prohibit	Not Applicable	Avg Context Length	Avg Reference #
Real	87	20	107	311.87	4.64
Synthetic	269	40	309	187.30	1.09

Table 1: Statistics of the evaluation datasets.

	Type	DP	CoT-auto	CoT-manual	Agent-ID	BM25-content	CI-ES-content
Llama3-instruct-8b	Real	77.57	79.43	72.89	86.44	**87.85**	85.98
	Synthetic	82.52	93.52	94.49	94.49	**95.46**	95.30
Qwen1.5-14b	Real	35.98	**87.38**	78.50	81.77	85.04	83.17
	Synthetic	48.86	**96.27**	95.46	94.26	95.46	94.98
Qwen2-7b	Real	48.13	68.69	63.55	71.02	67.75	**79.44**
	Synthetic	64.23	81.55	79.77	80.90	82.52	**88.67**
GLM-4-chat-9b	Real	64.95	70.09	73.83	77.10	**82.71**	76.63
	Synthetic	89.48	94.17	**95.30**	91.90	91.74	94.01
Mistral-v0.3-7b	Real	60.28	64.01	63.55	**69.62**	69.15	**69.62**
	Synthetic	85.59	82.68	92.07	92.07	**92.23**	90.27
GPT-4-turbo-04-09	Real	86.91	74.76	88.31	89.25	**89.71**	86.91
Average	Real	62.30	74.06	73.43	79.20	**80.36**	80.29
	Synthetic	74.13	89.63	91.41	90.72	91.48	**92.64**

Table 2: The overall accuracy evaluation results. The accuracies are reported in%.

6.1 Overall Assessments

In this section, we evaluate the overall accuracy of the proposed six types of prompts over various LLMs. Table 2 reports the accuracies for both *real* and *synthetic* court cases and includes six prompts' average performance. According to the results, we summarize the following findings:

1) *Our proposed Privacy Checklist is effective in improving LLMs' in-context reasoning ability to decide privacy violations.* In general, *agent-ID*, *BM25-content* and *CI-ES-content* outperform *DP*, *CoT-auto* and *CoT-manual*, leading to around 6% - 18% accuracy improvements. These improvements can be attributed to the retrieved sub-rules from our *Privacy Checklist*.

2) In terms of prompting tricks, we observe that *manually crafted prompts based on contextual integrity theory leads to improved LLM performance.* In our experiments, *CoT-manual* generally outperforms *DP* and *CoT-auto*. Despite *CoT-auto* also exploits LLMs' planning abilities to decompose the query into step-by-step plans, its automatic planning prompts still underperform *CoT-manual*'s expert taxonomies by approximately 5%. This finding suggests that contextual integrity theory is beneficial for LLMs to reason about privacy violations.

3) *LLMs perform better at synthetic data rather than real court cases.* Even though the synthetic data is augmented via GPT-4, all open-source LLMs lead to significant accuracy improvements over all six prompting methods. We manually inspect the synthetic samples and observe that LLMs tend to simplify the context to overfit the given instructions. The context information of most synthetic data shifts to naive and unrealistic narratives. Consequently, the distribution of synthetic data is dominated by easy-to-solve questions. Such distribution shift is also observed on LLM-augmented math problems (Tong et al., 2024).

4) *LLMs still need improvements to serve as responsible judges.* According to Table 2's results, all LLMs struggle to reach 90% accuracy for real court cases. Under the same prompt template, the results are highly influenced by LLMs' inherent reasoning and context-understanding ability. Powerful foundation models, such as GPT-4, yield the best performance for real court cases.

6.2 Inspections on Class-level Performance

In addition to evaluations of the overall accuracies, we also study the precision, recall and F1 within each class. Without loss of generality, we examine the detailed performance of one representative LLM, Llama-3-instruct-8b, for real court cases over the six prompts. Its precision, recall and F1 for each class are reported in Table 3. Our results imply the following observations:

5) *LLMs are capable of determining applicability and perform badly at prohibited cases.* Cases that are not applicable achieve over 90% F1 score for all prompts. In contrast, permitted cases have an average F1 of 77.24% and prohibited cases only have an average F1 of 45.59%. These results sug-

	Permit			Prohibit			Not Applicable		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
DP	87.67	73.56	80.00	45.83	55.00	50.00	**97.84**	85.04	91.00
CoT-auto	86.36	65.51	74.50	27.27	60.00	37.50	97.11	94.39	95.73
CoT-manual	84.09	42.52	56.48	22.41	**65.00**	33.33	95.49	99.06	97.24
Agent-ID	89.47	78.16	83.43	52.63	50.00	51.28	90.67	**100.00**	95.11
BM25-content	87.05	**85.05**	**86.04**	**60.00**	45.00	**51.42**	92.92	98.13	**95.45**
CI-ES-content	**91.66**	75.86	83.01	45.83	55.00	50.00	90.67	100.00	95.11
Average	87.72	70.11	77.24	42.33	55.00	45.59	94.12	96.10	94.94

Table 3: The detailed investigation of Llama-3's performance over 3 classes on the real court cases. Results are reported in %.

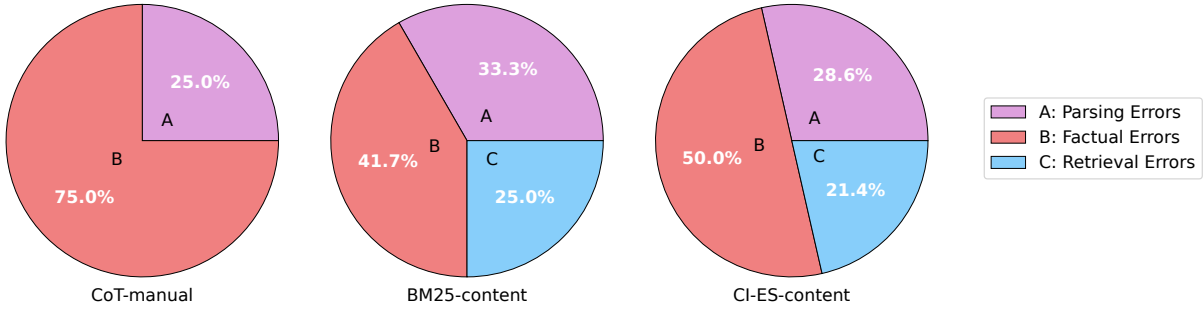

Figure 3: Manual investigations on GPT-4's errors for prohibited cases.

gest that all LLMs are impotent and biased judges on prohibited cases even if their contexts are given.

6) *CoT prompting only improves LLMs' performance on applicability, while our checklist helps LLMs to make correct judgments on permitted cases.* Though both CoT prompts lead to improved overall performance, such improvements mainly come from increased F1 on applicability. On the other hand, their F1 scores on both permitted and prohibited cases drop significantly. CoT prompting fails to enhance LLMs to determine compliance for applicable cases. Instead, the RAG-based pipelines not only achieve comparable performance on applicability as CoT prompting, but also improve LLMs' performance on permitted cases.

6.3 Error Analysis via Human Evaluations

In this section, we conduct a qualitative analysis to manually inspect LLMs' erroneous predictions. Two authors and one law student examine 60 prohibited cases' logs of *CoT-manual*, *BM25-content* and *CI-ES-content* generated by GPT-4. Upon investigation, we summarize three types of errors. The first is the parsing error, where our parsers fail to capture the correct predictions. The second is factual errors, which include hallucinations and misinterpretations of the context. The last is retrieval errors whose retrieved rules are irrelevant and misleading. Figure 3 depicts the error distributions for the three RAG-based prompts. Since *CoT-manual* is not based on the RAG pipeline, so its retrieval error is 0%.

According to the pie charts, though GPT-4 is one powerful foundation model, its factual errors still constitute the most significant portion. Such factual errors are commonly caused by hallucinations where LLMs reason on non-existing assumptions. On the other hand, though prompting with our *Privacy Checklist* can reduce the rate of factual errors, more than 20% of errors are caused by the irrelevant retrieved sub-rules. These irrelevant sub-rules may further misguide LLMs in fitting their criteria and therefore generate far-fetched explanations. This observation implies that LLMs' performance can be further improved by reducing hallucinations and enhancing the retriever's functionality.

7 Conclusions and Future Works

In this paper, we follow the spirit of Contextual Integrity Theory to transform privacy-related issues into reasoning problems. Unlike prior attempts which manage to build rigorous expert systems based on logic languages, we exploit powerful LLMs to further formulate the reasoning problems to the in-context reasoning problems based on the retrieval-augmented generation (RAG) pipeline. To facilitate the retrieval process, we propose *Privacy Checklist*, a checklist that covers processed legal documents, annotated CI characteristics, and auxiliary knowledge about roles, private attributes and term definitions. We compare three RAG-based prompts with three baseline prompts to show that

our proposed *Privacy Checklist* can improve LLMs' privacy awareness for real and synthetic court cases. Our future works can be divided into two folds. First, we plan to integrate most mainstream privacy regulations as well as AI safety standards into our *Privacy Checklist*. Second, we will work on improving the retrieval performance to eliminate the gap between common sense knowledge and legal terminologies. After getting over the two parts, we may no longer need the privacy-utility trade-off for model training and inference. Instead, we may simply apply our checklist as a detector for safeguarding models' inputs and outputs and concentrate on models' utility improvements.

Acknowledgements

The authors of this paper were supported by the ITSP Platform Research Project (ITS/189/23FP) from ITC of Hong Kong SAR, China, and the AoE (AoE/E-601/24-N), the RIF (R6021-20) and the GRF (16211520 and 16205322) from RGC of Hong Kong SAR, China.

Limitations

When we apply our *Privacy Checklist* to LLMs to evaluate compliance with a given court case, a major limitation is that LLMs cannot perform well on prohibited cases, even for GPT-4 models. As mentioned in the experimental results of Table 3, LLMs are impotent and biased judges on prohibited cases. After manually inspecting all prohibited cases, we find that LLMs are likely to assume that conditions not presented in the context, but required by the retrieved regulations, are satisfied. Such hypotheses cause LLMs to hallucinate and further conclude that a positive norm is satisfied for prohibited cases. This limitation suggests current LLMs are not ready yet to be legal agents. Instead, we should further tune LLMs to align with given contexts without unnecessary "if statements" to use LLMs to detect privacy violations.

Additionally, since LLMs with 7-14 billion parameters usually have context lengths of no more than 8k, we cannot feed few-shot demonstrations into these LLMs to evaluate the few-shot performance.

Ethical Considerations

We declare that all authors of this paper acknowledge the *ACM Code of Ethics* and honor the ACL code of conduct. Our work claims that simply studying leakage of PII patterns may not be enough to align with people's actual concerns. Instead, we consider the context-centric approach to determine privacy violations grounded on established privacy norms and standards. We believe that our checklist will become a new paradigm for studying privacy issues as LLMs are consistently improving their reasoning abilities.

Checklist Construction. During the construction process of the *Privacy Checklist*, we parse the HIPAA document tree from the official website of the Code of Federal Regulations following their granted access rules. In terms of expert verification steps, all manual annotations are initially done by two of the authors and another student with a law degree who can provide professional suggestions.

Evluation Data. Both synthetic and real court cases used in our evaluation are publicly available from GoldCoin's official GitHub implementation[3] under the Apache-2.0 license. For the real court case collection procedure, GoldCoin follows the official usage and access rules of the Caselaw Access Project[4] during downloading relevant cases.

[3] https://github.com/HKUST-KnowComp/GoldCoin
[4] https://case.law/about/#usage-access

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. CCS '16, New York, NY, USA. Association for Computing Machinery.

AI@Meta. 2024. Llama 3 model card.

A. Barth, A. Datta, J.C. Mitchell, and H. Nissenbaum. 2006. Privacy and contextual integrity: framework and applications. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*, pages 15 pp.–198.

Sebastian Benthall, Seda Gürses, and Helen Nissenbaum. 2017. Contextual integrity through the lens of computer science. 2(1):1–69.

Gregory Buehrer, Bruce W. Weide, and Paolo A. G. Sivilotti. 2005. Using parse tree validation to prevent sql injection attacks. In *Proceedings of the 5th International Workshop on Software Engineering and Middleware*, SEM '05, page 106–113, New York, NY, USA. Association for Computing Machinery.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *Proceedings of USENIX Security Symposium*, pages 2633–2650.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *Preprint*, arXiv:2306.16092.

Edoardo Debenedetti, Giorgio Severi, Nicholas Carlini, Christopher A Choquette-Choo, Matthew Jagielski, Milad Nasr, Eric Wallace, and Florian Tramèr. 2023. Privacy side channels in machine learning systems. *arXiv preprint arXiv:2309.05610*.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. Jailbreaker: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*.

Wei Fan, Haoran Li, Zheye Deng, Weiqi Wang, and Yangqiu Song. 2024. Goldcoin: Grounding large language models in privacy laws via contextual integrity theory. *arXiv preprint arXiv:2406.11149*.

Jose Fonseca, Marco Vieira, and Henrique Madeira. 2009. Vulnerability & attack injection for web applications. In *2009 IEEE/IFIP International Conference on Dependable Systems & Networks*, pages 93–102.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Kang Gu, Ehsanul Kabir, Neha Ramsurrun, Soroush Vosoughi, and Shagufta Mehnaz. 2023. Towards sentence level inference attack against pre-trained language models. *Proc. Priv. Enhancing Technol.*, 2023:62–78.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Re, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John J Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sanonda Datta Gupta and Torsten Hahmann. 2023. Oppo: An ontology for describing fine-grained data practices in privacy policies of online social networks. *arXiv preprint arXiv:2309.15971*.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *Preprint*, arXiv:2305.15062.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023a. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*.

Haoran Li, Yangqiu Song, and Lixin Fan. 2022. You don't know my favorite color: Preventing dialogue representations from revealing speakers' private personas. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5858–5870, Seattle, United States. Association for Computational Linguistics.

Haoran Li, Mingshi Xu, and Yangqiu Song. 2023b. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14022–14040, Toronto, Canada. Association for Computational Linguistics.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yanhong Zheng, and Yang Liu. 2023. Prompt injection

attack against llm-integrated applications. *ArXiv*, abs/2306.05499.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Gopi Nath Nayak and Shefalika Ghosh Samaddar. 2010. Different flavours of man-in-the-middle attack, consequences and feasible solutions. In *2010 3rd International Conference on Computer Science and Information Technology*, volume 5, pages 491–495.

Roger M. Needham and Michael D. Schroeder. 1978. Using encryption for authentication in large networks of computers. *Commun. ACM*, 21(12):993–999.

Helen Nissenbaum. 2010. Privacy in context: Technology, policy, and the integrity of social life. *Bibliovault OAI Repository, the University of Chicago Press*.

Harshvardhan J. Pandit, Axel Polleres, Bert Bos, Rob Brennan, Bud Bruegger, Fajar J. Ekaputra, Javier D. Fernández, Roghaiyeh Gachpaz Hamed, Elmar Kiesling, Mark Lizar, Eva Schlehahn, Simon Steyskal, and Rigo Wenning. 2019. Creating a vocabulary for data privacy. In *On the Move to Meaningful Internet Systems: OTM 2019 Conferences*, pages 714–730, Cham. Springer International Publishing.

Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *arXiv preprint*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP 2019*. Association for Computational Linguistics.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

R. Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2016. Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the exploitability of instruction tuning. *arXiv preprint arXiv:2306.17194*.

Yan Shvartzshnaider, Schrasing Tong, Thomas Wies, Paula Kift, Helen Nissenbaum, Lakshminarayanan Subramanian, and Prateek Mittal. 2016. Learning privacy expectations by crowdsourcing contextual informational norms. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 4(1):209–218.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 377–390, New York, NY, USA. Association for Computing Machinery.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Proceedings of the NeurIPS 2020*.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. 2023. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Wikipedia. IRAC. [Online; accessed 15-October-2024].

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-Yang Chen, Kexin Yang, Mei Li, Min Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yunyang Wan, Yunfei Chu, Zeyu Cui, Zhenru Zhang, and Zhi-Wei Fan. 2024. Qwen2 technical report.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model. *Preprint*, arXiv:2406.04614.

A Potential Impacts of Checklist

In this section, we discuss the potential impacts of our proposed *Privacy Checklist*.

A.1 Connections with Legal Reasoning

American legal scholars commonly decompose legal reasoning into four sequential stages, including Issue, Rule, Application, and Conclusion (IRAC) (Guha et al., 2023; Wikipedia). "Issue" relates to issue-spotting to identify the legal issues in the given context. "Rule" aims to recall relevant legal rules for identified issues. Then, lawyers apply these rules ("Application") to analyze the rule applicability. Lastly, lawyers reach their conclusions after the application stage. Our checklist follows a similar spirit of IRAC to enable retrieval augmented generations for LLMs. For the retrieval process, we actually complete the Issue and Rule stages. Then, we leverage LLMs to finally handle the Application and Conclusion stages. Unlike existing legal reasoning tasks that are confined to evaluating partial stages of IRAC (Guha et al., 2023; Pipitone and Alami, 2024), our checklist enables LLMs to complete all 4 stages consecutively, which can be regarded as the ultimate objective of legal reasoning.

A.2 Broad Use Cases

Our checklist offers new solutions for the following 3 use cases.

1) A generative privacy breaches classifier. Current safety alignment methods lead to over-defense problems and safety data covers approximately 20% - 50% alignment data. Our checklist can be used to train an external guard to identify privacy-intruding model prompts and outputs to avoid potential harm. Then, the internal model only needs to be aligned for utility improvements.

2) A general privacy consultant. In addition, our checklist can serve as a privacy consultant for the public interest. Users are able to receive in-time responses about their privacy concerns.

3) A detector for conflicting legal documents. In the future, after we cover more regulations, our checklist can even be used to enumerate all possible roles and attributes to help legal scholars identify unnoticed conflicts between state laws and federal laws to determine preemption.

B Experimental Details

Computational resources. During our experiment, we use 2 NVIDIA RTX 6000 to run our codes, and it takes GPU hours around 2 weeks to complete all experiments.

Generation details. For open-source models, we generate the models' responses with the recommended configurations in their model cards. For close-source models, we use their official APIs to obtain the responses with temperature = 0 and top_p = 0.95 to ensure reproducibility.

Prompt templates for evaluation. In this paragraph, we list all prompt templates used to evaluate LLMs' performance. The full prompts for the first three non-RAG methods including **DP**, **CoT-auto** and **CoT-manual** are shown in Table 7. The remaining three RAG-based methods, generally follow the retrieval, filter, and decision-making pipeline. For the retrieval stage, different retrieval methods are used to get relevant sub-rules. For the filter stage, the same prompt can be applied to all methods to verify the retrieved sub-rule is relevant to the given event. Finally, the decision-making stage follows previous chain-of-thought reasoning prompts by incorporating the selected reference regulations after the filter stage. Prompts used for **Agent-ID** are presented in Table 8. Those for **BM25-content** can be found in Table 9, while Table 10 displays the prompts used for **CI-ES-content**.

Licenses. For our evaluation data, we use data provided by GoldCoin's official GitHub implementation under the Apache-2.0 license. In terms of used models, we follow all the licenses to use their models for the research propose. For example, we follow Llama 3 Community License Agreement to use the Llama3-8b model for experiments.

C Privacy Leakage on ML Models

Recent studies on privacy issues can be categorized as training data leakage and inference data leakage. For training data leakage, private data may be unintentionally collected for training and extracted via training data extraction attacks (Carlini et al., 2021; Li et al., 2023a; Deng et al., 2023). In terms of inference data leakage, users' private data used during inference may also be recovered. For example, sensitive system-level prompts may be extracted via prompt injection attacks (Wan et al.,

2023; Perez and Ribeiro, 2022; Liu et al., 2023; Shu et al., 2023) and side-channel attacks (Debenedetti et al., 2023). Moreover, information leakage (Gu et al., 2023; Li et al., 2022; Song and Raghunathan, 2020) may also occur on models' hidden representation to recover sensitive information.

D CI Characteristics Extraction Details

In this section, we give the detailed prompts used to extract CI characteristics for each leaf node in our parsed document tree $\mathcal{T}$. Our full prompt is shown in Table 5. We convert the CI characteristics extraction steps into a series of questions. Then, we use regular regular expressions to parse answers for each question. We re-run the prompt whenever we encounter a pattern-matching failure.

E Comparison with GoldCoin

In this section, we compare our checklist with GoldCoin (Fan et al., 2024) in the following aspects to show our contributions.

E.1 Usability

Current best-performing LLMs are either close-sourced with only commercial API accesses or open-sourced with giant model scales (i.e., 405B parameters) that are hard to fine-tune. Consequently, it is costly to apply GoldCoin to a wide range of models. Instead, our *Privacy Checklist* offers plug-in solutions to enhance LLMs on privacy compliance tasks with better usability.

E.2 Downstream Performance

Since GoldCoin and our checklist evaluate the performance of different LLMs with different setups (GoldCoin decomposes the three-way classification into two binary classification tasks), we cannot make a fair comparison.

We follow the setting of the binary compliance task in GoldCoin to compare our RAG-based methods with GoldCoin in Table 4 on open-sourced LLMs. According to the overall Ma-F1 score, our BM25-content generally outperforms GoldCoin with the advanced LLMs.

In terms of close-sourced LLMs such as GPT-4, GoldCoin cannot fine-tune these LLMs to improve model performance. Instead, our checklist enables RAG-based various methods to further improve close-sourced LLMs' performance, as shown in Table 2.

F Comparison between Synthetic and Real Data

In this section, we provide quantitative and manual analyses to illustrate why our checklist performs worse on the real data.

For the quantitative study, as shown in Table 1, we compare the average word numbers that a case contains (Avg Context Length) and average reference numbers that a given case is relevant to (Avg Reference #). The reference refers to the applicable norms. According to the result, it is obvious that real court cases have longer context lengths and reference numbers than synthetic cases, which suggests that real court cases are more complex.

For the manual evaluation, we randomly selected 50 samples from both synthetic and real data, including permitted and prohibited cases. For each pair of (real, synthetic) samples, we then analyzed and compared their contexts based on three key factors: complexity, authenticity, and consistency. Complexity assesses which vignette appears more complicated or detailed. Authenticity evaluates which story feels more realistic or genuine. Consistency determines which story maintains better internal coherence. We invite two postgraduate students with law degrees to consider which context is better regarding the three factors for the 50 sampled pairs and report the averaged results. As shown in Table 6, the results suggest that real court cases are much more complex, authentic, and consistent than synthetic cases. Even though synthetic cases look plausible, real cases have much more detailed and complex backgrounds. Moreover, real cases do not use the exact terminologies used for the HIPAA. Hence, real cases are more difficult to handle for LLMs.

Method	Models	Permit			Forbid			All
		Prec	Rec	F1	Prec	Rec	F1	Ma-F1
GoldCoin	MPT-7B	86.49	73.56	79.50	30.30	50.00	37.74	58.62
	Llama2-7B	84.21	91.95	**87.91**	41.67	25.00	31.25	59.58
	Mistral-7B	90.67	78.16	83.95	40.62	65.00	50.00	66.98
	Llama2-13B	87.80	82.76	85.21	40.00	50.00	44.44	64.83
Agent-ID	Llama3-8B	89.47	78.16	83.43	52.63	50.00	51.28	67.36
BM25-content	Llama3-8B	87.05	85.05	86.04	60.00	45.00	**51.42**	**68.73**
CI-ES-content	Llama3-8B	91.66	75.86	83.01	45.83	55.00	50.00	66.50

Table 4: Performance comparison between GoldCoin and our RAG-based methods. Ma-F1 denotes the macro-F1 score.

As a legal expert specializing in the HIPAA Privacy Rule, your task is to read a specific paragraph of the regulation <Regulation ID>:
<Regulation Content>
Now complete the following questions one by one:

Q1. ("Prohibit", "Permit" or "General Definition") Ascertain whether the regulation <Regulation ID> pertains to scenes that are:
A. Prohibit by law
B. Permit by law
C. General Definition

Q2. (Identify the stakeholders related to regulation <Regulation ID>) Identify the stakeholders related to the regulation <Regulation ID>. Your response must include the following seven characteristics about the flow of private information: [Sender, Sender Role, Recipient, Recipient Role, Subject, Subject Role, Information Type, Consent Form, Purpose]. Answer 'None' if no information about characteristics is present.
The "Sender," "Recipient," and "Subject" fields indicate the sender, recipient, and the data subject during information transmission.
The "Sender Role", "Recipient Role" and "Subject Role" fields indicate the role of the sender, recipient and subject (e.g., doctor, patient).
The "Information Type" field defines what kind of information would be passed, such as name or location.
The "Consent Form" field indicates whether the sender has obtained consent from the subject to send the message. If consent is required, you should answer "Consent" for a flexible requirement or "Authorization" for a formal and mandatory process required by the context. If consent is not related to the context, you should answer "None."
The "Purpose" field indicates the purpose of the mentioned information transmission, such as treatment, payment, or health care operations.

Q3: Are the Sender and Subject the same person?
A. Yes
B. No
C. Not Sure

Q4: Is Recipient and Subject the same person?
A. Yes
B. No
C. Not Sure

Q5. (Identify the relation between <Regulation ID> other sub-parts referred to in the context) Identify the relation between <Regulation ID> and referred <Regulation Reference List>. For each reference in <Regulation Reference List>, according to the context given, determine if the reference is an exception of <Regulation ID> or support the argument of <Regulation ID>. Answer with "Exception" or "Support" for each reference.

Table 5: Our CI characteristics extraction prompt. Light blue texts inside each "<>" block denote a string variable.

	Complexity			Authenticity			Consistency		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Real	93.00	3.00	4.00	85.00	2.00	13.00	57.00	40.00	3.00
Synthetic	4.00	3.00	93.00	13.00	2.00	85.00	3.00	40.00	57.00

Table 6: Manual comparison between real and synthetic data. Results are reported in %.

DP
Determine the relationship between the provided event and HIPAA regulations. You have three options for your conclusion:
(A) The event is prohibited by HIPAA.
(B) The event is permitted by HIPAA.
(C) The event is not related to HIPAA.
Event: <event>
Output Format:
Choice: [A. Prohibited | B. Permitted | C. Not related]

CoT-auto
Determine the relationship between the provided event and HIPAA regulations. You have three options for your conclusion:
(A) The event is prohibited by HIPAA.
(B) The event is permitted by HIPAA.
(C) The event is not related to HIPAA.
Event: <event>
Let's approach this step by step:
Planning: Outline the steps (no more than 6) necessary to analyze the relationship of the event with HIPAA standards.
Execution: Implement these steps to gather information and assess the specific aspects of the event.
Decision Making: Based on the information collected and analyzed, determine if there was a HIPAA violation.
Output Format:
Plans: [List the steps planned to evaluate the event:]
plan 1 - ...
plan 2 - ...
....
plan N - ...
Execution: [Document the outcomes from executing your plans:]
plan 1 - ....
plan 2 - ....
....
plan N - ...
Choice: [A. Prohibited | B. Permitted | C. Not related]

CoT-manual
Determine the relationship between the provided event and HIPAA regulations. You have three options for your conclusion:
(A) The event is prohibited by HIPAA.
(B) The event is permitted by HIPAA.
(C) The event is not related to HIPAA.
Event: <event>
Let's approach this step by step:
1. Identify Relevant HIPAA Regulations: Start by researching and compiling the HIPAA regulations that could be relevant to the event, particularly focusing on privacy, security, and breach notification rules.
2. Examine the Nature of Information: Determine if the event involves any Protected Health Information (PHI), such as names, medical records, or other personal identifiers.
3. Evaluate HIPAA Applicability: Assess whether the event actually falls under the jurisdiction of HIPAA based on the nature of the entity involved and the type of information handled.
4. Assess Authorization and Consent: If HIPAA is applicable, verify if the recipient of the information was authorized and check for valid patient consent or another legal basis for sharing PHI.
5. Evaluate Security Measures: If applicable, review the security measures used during the information transmission to ensure they comply with HIPAA's security protocols.
Output Format:
Execution: [Document the outcomes from executing your plans:]
plan 1 - ....
plan 2 - ....
....
plan 5 - ....
Choice: [A. Prohibited | B. Permitted | C. Not related]

Table 7: Prompts used without retrieval augmented generation. Light blue texts inside each "<>" block denote a string variable.

LLMs as Knowledge Bases to Retrieve Regulation IDs.
Read the event described below and generate the applicable HIPAA regulations (no more than {generated_num}). This regulation will assist in determining if the event violates HIPAA security principles in a downstream task.
Event: <event>
Let's complete it step by step:
1. Review the Event Details: Understand the specifics of the event, including the type of information sent, the method of transmission, and the parties involved.
2. Identify Key HIPAA Concerns: Based on the event, identify potential concerns related to privacy, security, and breach notifications.
3. Retrieve Relevant Regulations: Consult the HIPAA regulatory text to find sections specifically addressing the identified concerns. Consider feedback to avoid repeating previously rejected regulations.
Output Format:
Execution: [Document the outcomes from executing the steps:]
1. - ...
2. - ...
...
4. - ...
Generated Related HIPAA Regulations(e.g. re"[0-9]+[̇0-9]+(\([0 - 9A-Za-zivx]+\))*" - xxxx):
1. Section Number - Section Content
...
N. Section Number - Section Content
Output:

Law Filter Template.
Given a regulation and an information sending or requesting event, identify if the given regulation is relevant to the event. The process involves the following steps:
1. **Understand the Event**: - Extract key details from the description of the event, such as the type of information being exchanged, the parties involved, and the context or domain (e.g., healthcare, finance, education).
2. **Analyze The Regulation**: - For given regulation, determine its scope and main focus by reading the summary or key sections. Identify the primary subject matter, applicable contexts, or targeted stakeholders.
3. **Relevance Matching**: - Compare the key details of the event with the main focus of the regulation. Consider: - Does the regulation explicitly address scenarios similar to the event? - Is the regulation intended for the context or domain of the event? - Are the parties involved in the event the type of entities the regulation aims to govern?
4. **Filtering Decision**: - If a regulation's focus aligns well with the event's details, please answer yes. - If there is little to no alignment, such as different contexts, unrelated subjects, or inappropriate scopes, please answer no.
Event: <event>
HIPAA Regulation: <candidates>
Note: It is possible that the event may be completely unrelated to the HIPAA regulations provided. In such cases, just answer NONE.
Question: Is the given HIPAA Regulation relevant with the given event?
Output Format: First answer yes or no, then explain the reason based on the given steps.

Decision-making Template.
Determine the relationship between the provided event and HIPAA regulations. You have three options for your conclusion:
(A) The event is prohibited by HIPAA.
(B) The event is permitted by HIPAA.
(C) The event is not related to HIPAA.
I will offer you some retrieved HIPAA regulations for reference(Not 100% correct.) Consider the specifics of the event, offered sections of HIPAA regulations.
Event: <event>
HIPAA Regulations Reference: <reference_regulations>
Let's complete it step by step:
1. Understand the Event: Read the description of the event carefully to know exactly what happened.
2. Look Up HIPAA Rules: Get the HIPAA regulations that are provided and find the parts that might relate to the event.
3. Check for Key Points: Focus on important details of the event like the kind of information involved, who is handling it, and how it's being shared or used.
4. Compare the Event with the Rules: See how the details of the event stack up against the HIPAA rules to find any matches or issues.
Output Format:
Execution: [Document the outcomes from executing the steps:]
1. - ...
2. - ...
...
4. - ...
Choice: [A. Prohibited | B. Permitted | C. Not related]

Table 8: Prompts used for **Agent-ID**. Light blue texts inside each "<>" block denote a string variable.

LLM Explanation for BM25 Similarity.
I will provide you with an event concerning the delivery of information. Your task is to generate content related to this event by applying your knowledge of HIPAA regulations.
To ensure the content is relevant and accurate, follow these steps:
1. Understand the Event: Clearly define and understand the specifics of the event. Identify the key players involved, the type of information being handled, and the context in which it is being delivered.
2. Apply HIPAA Knowledge: Utilize your understanding of HIPAA regulations, focusing on privacy, security, and the minimum necessary information principles. Ensure that your content addresses these aspects in the context of the event.
Event Details: <event>
Output Format:
Execution:
1. Identify the key players, type of information, and context.
2. Apply relevant HIPAA principles to the event.
Generated HIPAA Content:
1. HIPAA Privacy Rule: ...
2. HIPAA Security Rule: ...
3. Minimum Necessary Standard:...
References:
List the specific HIPAA regulations you consulted to generate the content.

Law Filter Template.
Given a regulation and an information sending or requesting event, identify if the given regulation is relevant to the event. The process involves the following steps:
1. **Understand the Event**: - Extract key details from the description of the event, such as the type of information being exchanged, the parties involved, and the context or domain (e.g., healthcare, finance, education).
2. **Analyze The Regulation**: - For given regulation, determine its scope and main focus by reading the summary or key sections. Identify the primary subject matter, applicable contexts, or targeted stakeholders.
3. **Relevance Matching**: - Compare the key details of the event with the main focus of the regulation. Consider: - Does the regulation explicitly address scenarios similar to the event? - Is the regulation intended for the context or domain of the event? - Are the parties involved in the event the type of entities the regulation aims to govern?
4. **Filtering Decision**: - If a regulation's focus aligns well with the event's details, please answer yes. - If there is little to no alignment, such as different contexts, unrelated subjects, or inappropriate scopes, please answer no.
Event:<event>
HIPAA Regulation: <candidates>
Note: It is possible that the event may be completely unrelated to the HIPAA regulations provided. In such cases, just answer NONE.
Question: Is the given HIPAA Regulation relevant with the given event?
Output Format: First answer yes or no, then explain the reason based on the given steps.

Decision-making Template.
Assess the relationship between the provided event and HIPAA regulations, and select one of the following conclusions:
(A) The event is prohibited by HIPAA.
(B) The event is permitted by HIPAA.
(C) The event is not related to HIPAA.
You will be given some sections of HIPAA regulations for reference (Note: these references may not be 100% accurate). Your task is to consider the specifics of the event in light of these sections. Remember that the provided references are not exhaustive, and the presence of a regulation does not automatically imply relevance or violation.
Event: <event>
HIPAA Regulations Reference: <reference_regulations>
Let's complete it step by step:
1. Analyze the specific details of the event by identifying who is involved (sender and recipient), what information is being sent or requested, and for what purpose.
2. Compare key elements of the event with HIPAA rules, identifying if they involve the use, disclosure, or protection of Protected Health Information (PHI) as defined by HIPAA.
3. Evaluate the provided HIPAA regulation excerpts to see if they directly relate to the event.
4. Consider if there are other HIPAA rules not mentioned in the excerpts that might apply.
5. Conclude based on the comprehensive analysis whether the event is in compliance, in violation, or unrelated to HIPAA.
Output Format:
Execution: [Document the outcomes from executing each step]:
1. - ...
2. - ...
...
5. - ...
Choice: [A. Prohibited | B. Permitted | C. Not related]

Table 9: Prompts used for **BM25-content**. Light blue texts inside each "<>" block denote a string variable.

Law Filter Template.
Given a regulation and an information sending or requesting event, identify if the given regulation is relevant to the event. The process involves the following steps:
1. **Understand the Event**: - Extract key details from the description of the event, such as the type of information being exchanged, the parties involved, and the context or domain (e.g., healthcare, finance, education).
2. **Analyze The Regulation**: - For given regulation, determine its scope and main focus by reading the summary or key sections. Identify the primary subject matter, applicable contexts, or targeted stakeholders.
3. **Relevance Matching**: - Compare the key details of the event with the main focus of the regulation. Consider: - Does the regulation explicitly address scenarios similar to the event? - Is the regulation intended for the context or domain of the event? - Are the parties involved in the event the type of entities the regulation aims to govern?
4. **Filtering Decision**: - If a regulation's focus aligns well with the event's details, please answer yes. - If there is little to no alignment, such as different contexts, unrelated subjects, or inappropriate scopes, please answer no.
Event:<event>
HIPAA Regulation: <candidates>
Note: It is possible that the event may be completely unrelated to the HIPAA regulations provided. In such cases, just answer NONE.
Question: Is the given HIPAA Regulation relevant with the given event?
Output Format: First answer yes or no, then explain the reason based on the given steps.

Decision-making Template.
Assess the relationship between the provided event and HIPAA regulations, and select one of the following conclusions:
(A) The event is prohibited by HIPAA.
(B) The event is permitted by HIPAA.
(C) The event is not related to HIPAA.
You will be given some sections of HIPAA regulations for reference (Note: these references may not be 100% accurate). Your task is to consider the specifics of the event in light of these sections. Remember that the provided references are not exhaustive, and the presence of a regulation does not automatically imply relevance or violation.
Event: <event>
HIPAA Regulations Reference: <reference_regulations>
Let's complete it step by step:
1. Analyze the specific details of the event by identifying who is involved (sender and recipient), what information is being sent or requested, and for what purpose.
2. Compare key elements of the event with HIPAA rules, identifying if they involve the use, disclosure, or protection of Protected Health Information (PHI) as defined by HIPAA.
3. Evaluate the provided HIPAA regulation excerpts to see if they directly relate to the event.
4. Consider if there are other HIPAA rules not mentioned in the excerpts that might apply.
5. Conclude based on the comprehensive analysis whether the event is in compliance, in violation, or unrelated to HIPAA.
Output Format:
Execution: [Document the outcomes from executing each step]:
1. - ...
2. - ...
...
5. - ...
Choice: [A. Prohibited | B. Permitted | C. Not related]

Table 10: Prompts used for **CI-ES-content**. Light blue texts inside each "<>" block denote a string variable.