GOLDCOIN: Grounding Large Language Models in Privacy Laws via Contextual Integrity Theory

Anonymous ACL submission

Abstract

Privacy issues arise prominently during the inappropriate transmission of information between entities. Existing research primarily studies privacy by exploring various privacy attacks, defenses, and evaluations within narrowly predefined patterns, while neglecting that privacy is not an isolated, context-free concept limited to traditionally sensitive data (*e.g.*, social security numbers), but intertwined with intricate social contexts that complicate the identification and analysis of potential privacy violations. The advent of Large Language Models (LLMs) offers unprecedented opportunities for incorporating the nuanced scenarios outlined in privacy laws to tackle these complex privacy issues. However, the scarcity of open-source relevant case studies restricts the efficiency of LLMs in aligning with specific legal statutes. To address this challenge, we introduce a novel framework, GOLDCOIN, designed to efficiently ground LLMs in privacy laws for judicial assessing privacy violations. Our framework leverages the theory of *contextual integrity* as a bridge, creating numerous synthetic scenarios grounded in relevant privacy statutes (e.g., HIPAA), to assist LLMs in comprehending the complex contexts for identifying privacy risks in the real world. Extensive experimental results demonstrate that GOLDCOIN markedly enhances LLMs' capabilities in recognizing privacy risks across real court cases, surpassing the baselines on different judicial tasks.

1 Introduction

Privacy violations happen through improper information transmission, including the disclosure of personally identifiable information, inappropriate data collection, and unauthorized access, all of which contradict societal expectations (Martin and Nissenbaum, 2016) and legal statutes such as HIPAA (Act, 1996), COPPA (Aftab and Savitt, 1999), and GDPR (Voigt and Von dem Bussche, 2017). In the past few decades, current research has mainly focused on exploring privacy violations in limited pre-defined patterns or manually annotated rules, such as RBAC (Sandhu, 1998; Kuhn et al., 2010), EPAL (Ashley et al., 2003, 2002), thereby diminishing the capacity to detecting privacy risks across diverse social contexts.

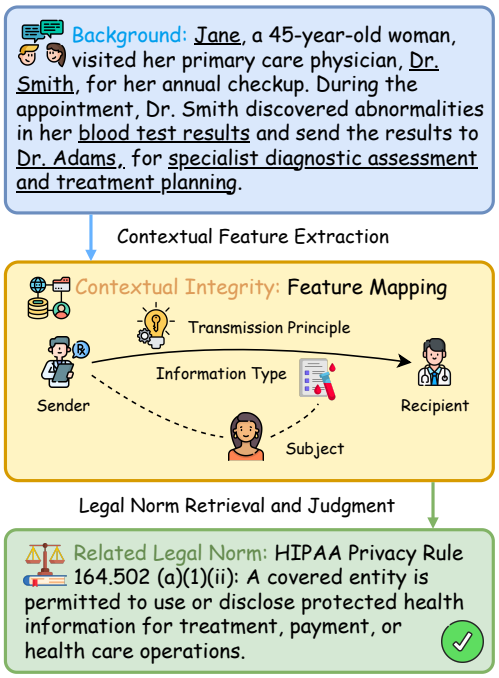

Figure 1: An overview of how our proposed GOLDCOIN bridge case background and legal norms through contextual integrity theory (Nissenbaum, 2004).

Intuitively, we consider applying the wealth of real-world scenarios contained in legal statutes and case law to address the limitation. However, converting legislation into an actionable framework remains a significant challenge. Previous efforts have involved translating legislation into logical languages (Lam et al., 2009; DeYoung et al., 2010; Robaldo et al., 2020), yet this method heavily relies on expert annotation and struggles to adapt to legislative changes or scale across different privacy laws. The recent emergence of LLMs (OpenAI,

2022; Touvron et al., 2023b; Anthropic, 2024), has introduced new potential for addressing the problem. Specifically, legal LLMs like LawGPT (Zhou et al., 2024), Lawyer LLaMA (Huang et al., 2023), ChatLaw (Cui et al., 2023) have all leveraged the vast existing statutes and cases to assist public in general legal tasks.

Nonetheless, aligning LLMs with specific privacy laws is a non-trivial task. The scarcity of open-source public court cases makes it challenging to ensure that the datasets used in model training are comprehensive enough to encompass all aspects of the laws. This limitation significantly undermines the LLMs' ability to generalize to unfamiliar cases. Moreover, we observe unstable and limited improvements when training LLMs directly on statutory laws (Section 5.2), as court cases generally provide a richer source of practice-oriented information, such as factual backgrounds, judicial analyses, and judge opinions.

To fill these gaps, we introduce GOLDCOIN, a novel framework that **G**r**O**unds **L**arge Language Mo**D**els into Privacy Laws via **CO**ntextual **IN**tegrity, which is a theory proposed by Nissenbaum (2004) to assess the appropriateness of privacy information flows. Within contextual integrity, privacy information flows are conceptualized as activities involving three relevant entities: the sender, the recipient, and the subject of the information. It argues that entities do not merely act as individuals in an undifferentiated social world (Barth et al., 2006), but rather as individuals playing various roles within specific contexts (*e.g.*, healthcare, education, employment). Within each distinct context, information flows are regulated by norms (*a.k.a.* regulations, legal clauses) that specify the types of the transmitted information and the transmission principles (*e.g.*, purpose, consent, belief). Then we can abstract privacy laws as the framework for determining the legality of information flow in diverse contexts, including entities, information type, and transmission principles. Each clause in privacy laws, such as `164.502(a)(1)(ii)` referenced in Figure 1, can be interpreted as a legal-grounded norm, either permitting or forbidding information transmission.

Based on this, GOLDCOIN combines the formalization of contextual integrity with concrete seed norms in privacy laws to generate the synthetic background stories by GPT-4 (Achiam et al., 2023). To ensure high-quality generation, we employ automatic filters to select cases that include essential features (*e.g.*, sender, recipient) in contextual integrity and are consistent with the seed norms. Additionally, we develop a diversity ranking mechanism to improve the semantic diversity of the case backgrounds, enhancing training robustness. Ultimately, our framework combines background contexts and seed norms to construct synthetic court cases tailored to specific privacy laws.

For evaluation, we develop the case dataset GOLDCOIN-HIPAA under the HIPAA Privacy Rule (Act, 1996), including a ground-truth benchmark sourced from the Caselaw Access Project (CAP)[1] (Chang et al., 2020), which collectes numerous real-world court cases in the United States. We experiment with several transformer-based LLMs by instruction-tuning them with GOLDCOIN. The evaluation results demonstrate that our synthetic dataset effectively aids LLMs in comprehending privacy laws. The models tuned with our framework show superior ability in identifying the applicability of HIPAA in real cases, surpassing other baselines by 8% to 23%. Meanwhile, these models show enhanced capabilities in detecting privacy risks, outperforming others by 8% to 18%. Moreover, human analysis and ablation studies confirm the efficacy of contextual integrity in case synthesis and the enhancements in data quality provided by the automatic filter and diversity ranking.

2 Related Work

2.1 Privacy and Contextual Integrity

To effectively ground language models into privacy laws for judgment in reality, we first introduce the contextual integrity theory (Nissenbaum, 2004) and propose a brief framework based on the existing works (Barth et al., 2006; Lam et al., 2009).

Roles, Information, and Transmission Principle Each information transmission inherently involves three main entities $\mathcal{P}$: the ***sender***, ***recipient***, and ***subject*** whose information is about. The roles of these entities are deeply contextual, as individuals participate in specific ***roles*** $\mathcal{R}$ tailored to distinct social contexts such as healthcare and commerce. Moreover, the ***information type*** associated with the subject is denoted as $\mathcal{T}$. ***Transmission principles***, represented as Ω, comprise specific constraints $\omega \in \Omega$ (*e.g.*, purpose, authorization) that regulate the information flow.

[1] https://case.law/

Expressing in Norms Applying contextual integrity to the privacy law $\mathcal{L}$, we can abstract each legal clause as a norm n, which governs the information flow between entities.

$$\begin{aligned} \textit{permitted}_{by} n^{+} &\iff (\mathcal{P}, \mathcal{R}) \wedge \mathcal{T} \wedge \Omega, \\ \textit{forbidden}_{by} n^{-} &\iff (\mathcal{P}, \mathcal{R}) \wedge \mathcal{T} \wedge \Omega, \end{aligned} \tag{1}$$

where a **`permit`** norm n^{+} allows an information transmission when satisfying conditions, and a **`forbid`** norm n^{-} prohibits it when aligning with the specified features. Further details and examples are provided in the Appendix A.

2.2 LLMs in Law

Recent advancements in legal LLMs, such as LawGPT (Nguyen, 2023; Zhou et al., 2024), Lawyer LLaMA (Huang et al., 2023; Touvron et al., 2023a) and SaulLm (Colombo et al., 2024) have shown significant improvements in a broad array of legal services, including judgment prediction (Yue et al., 2021a; Zhang et al., 2023), court view generation (Yue et al., 2021b), and question answering (Duan et al., 2019; Zhong et al., 2020). ChatLaw (Cui et al., 2023) specializes in processing Chinese legal queries, excelling in keyword extraction and court case similarity-matching. However, these LLMs often underperform in the privacy domain, where related training and evaluation datasets are limited and generally close-sourced.

2.3 LLMs for Instruction Generation

A series of works have explored the use of LLMs for data generation (Meng et al., 2023; Liu et al., 2022; Wang et al., 2021; Schick and Schütze, 2021). Recent studies have particularly concentrated on enhancing instruction generation (Honovich et al., 2022a; Zhou et al., 2023; Singh et al., 2023; Honovich et al., 2022b; Wang et al., 2023) to improve zero-shot (Ye et al., 2022) and few-shot (Brown et al., 2020; Wei et al., 2021) learning, abstraction reasoning (Wang et al., 2024) capabilities, as well as the instruction-following proficiency of LLMs. Inspired by it, our approach utilizes the strong generative capabilities of GPT-4 to address the case scarcity based on contextual integrity theory by instructing it to generate datasets and align LLMs with privacy laws for judicial.

3 Method

Legal judgments on privacy violations typically involve two tasks (Lam et al., 2009): (1) **Applicability**, assessing whether the privacy law $\mathcal{L}$ applies to the case background s; and (2) **Compliance**, determining if the transmission described in s compliant with $\mathcal{L}$. In this section, we introduce GOLDCOIN, which applies contextual integrity theory to generate synthetic cases. After postprocessing the instances, we instruction-tune LLMs and evaluate their performance of the above two tasks. The overview of our pipeline is shown in Figure 3.

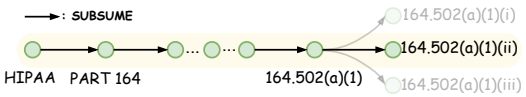

Figure 2: We concatenate all the content along the whole path from the leaf (`164.502(a)(1)(ii)`) to the root (`HIPAA`) node and refer to it as a norm, as illustrated in the norm part of Figure 8.

3.1 Legal Statute Preprocessing

To evaluate the effectiveness of GOLDCOIN, we apply it to the U.S. Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Initially, we dump the content of the HIPAA Privacy Rule from the official Code of Federal Regulations (CFR) website[2]. We then transform the textual data into a structured graph $\mathcal{G}$, comprising nodes $\mathcal{V}$ that represent sections and two types of relations $\mathcal{E}$. These relationships are identified as **`subsume`**, denoting hierarchical relationships (e.g., (`164.502(a)`, `subsume`, `164.502`)), and **`refer`**, indicating cross-references between sections (e.g., (`164.502(a)(1)(ii)`, `refer`, `164.504(b)`)). Each node consistes of a labeled identifier and the paragraph content. We start from each leaf node v_i^l and recursively identify all parent nodes $\{v_i^{l-1}, v_i^{l-2}, \ldots, v_i^0\}$ where v_i^0 is the root node. Then we aggregate their content, as depicted in Figure 2 and refer to each such path as a norm. These norms can be categorized into three types: **`permit`** n^{+}, **`forbid`** n^{-}, and others. The first two categories describe the permissions and prohibitions regarding information transmission under the law, while the last category contains general definitions, exceptions, and requirements. We leverage GPT-4 to classify and label these leaf norms and filter the permit and forbid norm for the subsequent generation steps. The examples of permit and forbid norms are shown in the upper part of Figure 8 and Figure 9, respectively. All the details of the HIPAA Privacy Rule and preprocessing are depicted in Appendix B.

[2] https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C

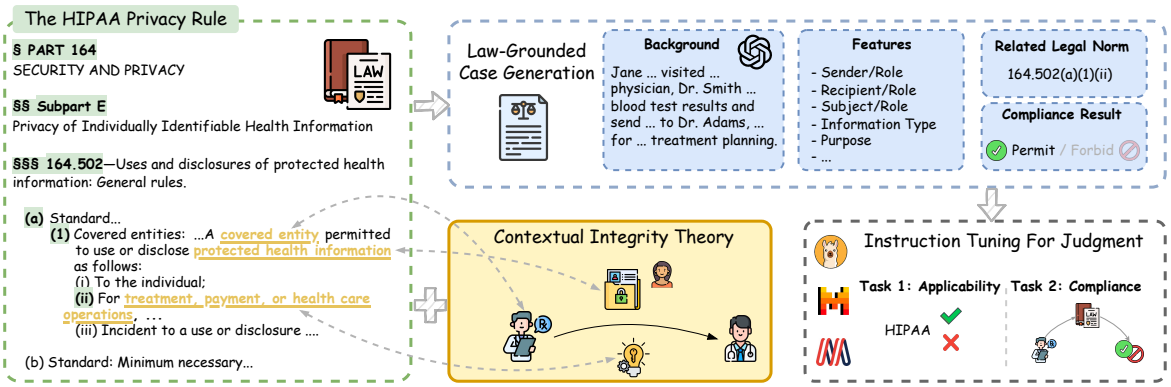

Figure 3: The overview of our GOLDCOIN framework. We use `164.502(a)(1)(ii)` as a seed norm to generate cases based on the contextual integrity theory and instruction-tune the models for downstream judicial tasks.

3.2 Law-Grounded Case Generation

After classification, we select the norms $\mathcal{N} = \{n_1, n_2, \ldots, n_m\}$, a filtered subset of $\mathcal{L}$, as seeds for case synthesis. Our objective is to generate the case set $\mathcal{K} = \{k_1, k_2, \ldots, k_m\}$, with each case k_i derived from n_i. In a synthetic case, four key elements are considered: *case background*, *contextual features*, *related norm*, and *conclusion*.

Instruction Compilation with Norm Given the seed norm n_i and the conclusion c_i which correspond to the norm type (*i.e.*, permit, forbid), we manually build the instructions combined with n_i for background generation. To ensure the generation of background narratives that align with n_i and preserve the integrity of the privacy information transmission context, we construct a detailed prompt (see Appendix C.1) that includes the description of the key features in contextual integrity, such as entities, roles, information type, and transmission principles.

Response Collection and Parsing To enhance the reliability of the model outputs, we sample several responses for each norm. Following the collection of GPT-4 outputs, we parse the responses and focus on the five components of k_i: (1) ***Background*** s_i, which is the background description of the information transmission. (2) ***Contextual features*** $\{(\mathcal{P}_i, \mathcal{R}_i), \mathcal{T}_i, \Omega_i\}$, which denotes the key features in the transmission context. (3) ***Norm*** n_i, which denotes the related legal clause (*e.g.*, `164.502(a)(1)(ii)`) to the generated background, (4) ***Applicability conclusion*** c_i^{appl}, which denotes whether the case applies to $\mathcal{L}$. (5) ***Compliance conclusion*** c_i^{comp}, which represents whether n_i permits or forbids the case.

3.3 Case Postprocessing

After collecting all cases, we implement several filters to ensure the consistency and quality of the selected cases.

Contextual Feature Integrity Filter By analyzing the key characteristics that are entailed in the generated cases, we observe that GPT-4 sometimes omits some main features in contextual integrity due to unstable instruction-following ability. To ensure the integrity of context in case background, we filter out all cases that lack any vital features in ***sender, sender role, recipient, recipient role, subject, subject role,*** and ***information type***.

Consistency Filter Each synthetic case is derived from a specific norm; however, the probabilistic variability of GPT-4 outputs may result in the related norm $\hat{n}$ of the synthetic case not aligning with the initial seed norm n. To improve the consistency of the cases, we filter out the cases that are not related to the given seed norm:

$$f_{\text{norm}}(n, \hat{n}) = \mathbb{1}(n = \hat{n}), \tag{2}$$

where f_{norm} denotes the compare function between the seed norm and the case norm. Moreover, we expect the model to generate cases applicable to $\mathcal{L}$ and its compliance c^{comp} (*i.e.*, permit or forbid) is consistent with the type of seed norm $n^{+/-}$:

$$\begin{aligned} f_{\text{conc}}(c^{appl}) &= \mathbb{1}(c^{appl} = applicable), \\ f_{\text{conc}}(n^{+/-}, c^{comp}) &= \mathbb{1}(n^{+/-} = c^{comp}), \end{aligned} \tag{3}$$

Then f_{conc} filters out all conclusion-inconsistent cases, ensuring that all cases apply to $\mathcal{L}$ and compliance with the seed norms.

Figure 4: Top 15 common sender roles (inner circle) and their top 10 recipient roles (outer circle).

Diversity Ranking To mitigate the reduction in semantic complexity and alignment robustness caused by similar features across different case backgrounds, we implement the methodology from prior studies (Wang et al., 2023, 2024) to promote background diversity. We calculate the ROUGE-L (Lin, 2004) similarity for each case background against others, ranking them accordingly. For each norm, we select the case with the highest ROUGE-L score to ensure optimal diversity.

3.4 Real Court Case Collection

To rigorously validate the grounding efficacy of GOLDCOIN, we retrieve all relevant real court cases featuring the "HIPAA Privacy Rule" from the Caselaw Access Project (CAP)[3], developed by Harvard Law School. These cases are systematically processed and distilled into the same format as synthetic cases, utilizing both GPT-4 and manual annotation. Additionally, as delineated in Section 3.2, our methodology focuses exclusively on constructing cases applicable to the HIPAA Privacy Rule, because privacy cases unrelated to HIPAA are readily available. To provide a negative training and testing set for the applicability task, we also curate a collection of cases that, while closely related to privacy violations, do not apply to HIPAA. The details of CAP are explained in Appendix D.

3.5 Instruction and Response Compilation

Using the case background as input, we manually build multi-step instructions (see Table 8(b) and Table 9(b)) with the Chain-of-Thought (CoT) prompting (Wei et al., 2022), for both applicability and compliance tasks as follows:

```
Instruction: <task-specific instruction>
Input: <case background>
```

We construct responses for applicability task following two steps: first, extracting context-specific features; second, determining HIPAA applicability:

```
Step1: <sender>, <recipient>, ...
Step2: Applicable/Not applicable
```

In the compliance task, we initially guide LLMs to extract features based on contextual integrity, focusing on principles of information transmission. Subsequently, we retrieve the relevant norm and finally determine whether the norm permits or prohibits the transmission as the format:

```
Step1: <sender>, <recipient>, ...
Step2: <norm id>, <norm content>
Step3: Permit/Forbid
```

4 GOLDCOIN-HIPAA Dataset Overview

In this section, we apply our framework to the HIPAA Privacy Rule as a case study and introduce the GOLDCOIN-HIPAA dataset. Statistics of the dataset are provided in Table 5.

4.1 GOLDCOIN Synthetic Cases

In HIPAA, we analyze 269 `permit` and 40 `forbid` norms, generating multiple cases for each norm. We filter and select the case with the highest ROUGE-L scores for each norm, ultimately collecting 309 cases. Comparisons between original and filtered cases in terms of ROUGE-L score distributions are shown in Figure 5, indicating the efficacy of diversity ranking. Moreover, we plot a sunburst chart to display the top 15 most common roles of ***sender*** and their top 10 ***recipients*** (Figure 4), and another chart to detail the common combination between information subjects and types (Figure 7).

Human Analysis To further investigate the quality, we enlist two experts to evaluate the synthetic cases using the criteria outlined in Table 1. The results confirm that all cases apply to HIPAA, with the majority being related to the seed norm.

4.2 CAP Real Court Cases

Due to reasons outlined in Section 3.4, we directly collect cases irrelevant to HIPAA as negative training examples. We select cases tagged with "Privacy Violation" using the "Most Relevant First" search function on the CAP website, gathering

[3] https://case.law/

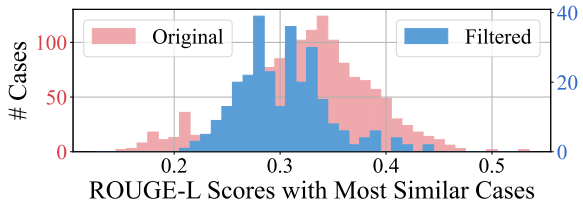

Figure 5: The ROUGE-L score distribution between the original and filtered cases.

Quality Review Question	Yes %
Does HIPAA apply to this case?	100.0%
Is the case strongly related to the seed norm?	99.35%
Is the compliance of the case correct?	99.03%
All fields are valid	98.38%

Table 1: Human analysis of synthetic case quality.

309 non-applicable cases for training. For evaluation, after a combined screening by GPT-4 and human experts, we identify 107 real court cases relevant to HIPAA, serving as the ground truth for the compliance task. Correspondingly, we also sample an equivalent number of HIPAA-irrelevant cases and combine them with the 107 cases to form the test set for the applicability task. Ultimately, we combine synthetic and real cases to create the GOLDCOIN-HIPAA dataset.

5 Experiment

In this section, we conduct extensive experiments to demonstrate the efficacy of GOLDCOIN in grounding LLMs into real-world privacy laws.

5.1 Experimental Settings

Datasets and Metrics As illustrated in Table 5, our framework generates 309 synthetic cases that either comply with or violate the HIPAA Privacy Rule (Act, 1996). Also, we collect 309 cases that do not apply to HIPAA. For evaluation, we collect 107 HIPAA-related and 107 unrelated real court cases from the CAP and calculate Accuracy (Acc) and Macro F1-score (Ma-F1) as metrics between predicted and ground truth conclusion.

Models We conduct instruction tuning on four open-source LLMs: MPT-7B-Chat-8k (MosaicML NLP Team, 2023), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Llama-2-7b-chat-hf and Llama-2-13b-chat-hf (Touvron et al., 2023b). These models all support at least 4k tokens content length and have superior instruction-following ability. Additionally, we evaluate our method against closed-source LLMs in zero-shot and few-shot settings, including models such as ChatGPT (gpt-3.5-turbo) (OpenAI, 2022) and GPT-4 (gpt-4) (Achiam et al., 2023; OpenAI, 2024), both with the version `2024-02-01` via Azure OpenAI API.

Baseline Methods We conduct comparative experiments against the following baselines to demonstrate the improvement introduced by GOLDCOIN. (1) Zero-shot: Given the background of cases, the LLMs should directly determine whether the case applies to HIPAA and violates HIPAA or not. (2) Law Recitation: No learning from cases, we tune the LLMs directly on the legal norm content. (3) Direct Prompt: Different from zero-shot, we instruction-tune the LLMs with vanilla prompts, where the responses are solely ("Applicable," "Not Applicable") or ("Permit," "Forbid"). The baseline prompts are shown in Appendix E.

5.2 Overall Performance

We present comprehensive results for two judicial tasks in Table 2, which includes the baseline methods and our GOLDCOIN. Besides, Figure 6 displays a comparison results with the GPT-series.

Applicability We first analyze the performance of four LLMs in determining the HIPAA applicability of real court cases sourced from CAP. Our results demonstrate that GOLDCOIN can align the LLMs with the comprehensive understanding of the HIPAA Privacy Rule, exceeding all baseline methods. Notably, MPT-7B, which performed near-random levels (Acc 50%, Ma-F1 50%), see substantial improvements with our method—accuracy and Macro F1-scores increase by 12.62% and 11.81%, respectively, compared to the zero-shot setting. Meanwhile, Mistral-7B and Llama2-13B tuned with our framework, achieve exceptional accuracy rates of 97.66% and 99.53%, respectively, even attaining 100% in "Not applicable" category (see Table 12), surpassing the performance of ChatGPT and GPT-4. We observe that MPT-7B, when trained exclusively with "Direct Prompt," exhibits only a limited improvement of 2.49% in Ma-F1. This underscores the integration of contextual features is crucial for decomposing and deeply understanding legal case topics. Additionally, our results indicate that merely continuing to train LLMs on legal statutes results in limited effectiveness and

Task	Method	MPT-7B		Llama2-7B		Mistral-7B		Llama2-13B	
		Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1
Applicability	Zero-shot	55.61	55.49	72.89	71.05	89.25	89.24	91.12	91.07
	Law Recitation	44.86	44.69	74.30	72.75	85.98	85.96	91.59	91.57
	Direct Prompt	63.55	57.97	89.25	89.13	95.33	95.32	94.39	94.39
	GOLDCOIN	68.22	67.30	94.39	94.39	97.66	97.66	**99.53**	**99.53**
Compliance	Zero-shot	46.73	40.75	56.07	47.14	50.47	49.02	65.42	56.71
	Law Recitation	39.25	32.43	42.99	41.69	53.27	43.23	68.22	59.79
	Direct Prompt	66.36	56.46	62.62	53.68	53.27	51.75	73.83	62.40
	GOLDCOIN	69.16	58.62	**79.44**	59.58	75.70	**66.98**	76.64	64.83

Table 2: Performance of four LLMs under three baselines and our GOLDCOIN, showing **Acc** and **Ma-F1** across both applicability and compliance tasks. We **bold** the best results and underline the second-best results in each task.

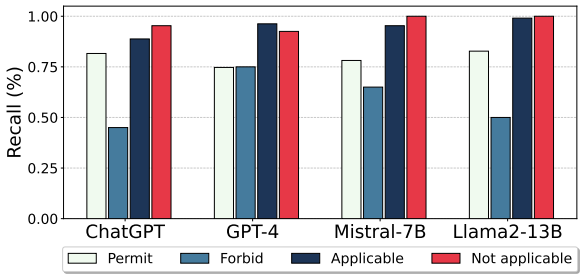

Figure 6: Comparative performance of GPT series models and our GoldCoin framework measured by **Recall** across all categories, with multi-step instructions.

Model	Applicability	Δ_{App}	Compliance	Δ_{Com}
Llama2-13B	**99.53**	-	**64.83**	-
◇ w/o Feature F	96.27	↓3.26	62.47	↓2.36
◇ w/o Norm F	97.59	↓1.94	61.34	↓3.49
◇ w/o Conclusion F	94.54	↓4.99	61.07	↓3.76
◇ w/o Diversity R	95.67	↓3.86	62.33	↓2.50
◇ w/o All Parts	93.01	↓6.52	60.11	↓4.72
Mistral-7B	**97.66**	-	**66.98**	-
◇ w/o Feature F	95.22	↓2.44	65.04	↓1.92
◇ w/o Norm F	95.98	↓1.68	63.34	↓3.62
◇ w/o Conclusion F	93.61	↓4.05	63.05	↓3.91
◇ w/o Diversity R	95.54	↓2.12	64.45	↓2.51
◇ w/o All Parts	91.77	↓5.89	61.91	↓5.05

Table 3: Ablation study for GOLDCOIN. Macro F1-scores are presented, with Δ indicating score changes.

even leads to diminished performance in determining applicability (*e.g.*, MPT-7B ↓10.8%).

Compliance Our GOLDCOIN framework introduces multi-step simulated trial instructions, effectively aligning LLMs with privacy law and enhancing their reasoning capabilities on compliance tasks. It significantly improved Macro-F1 scores across several models: MPT-7B (17.87%), Llama2-7B (12.45%), Mistral-7B (17.96%), and Llama2-13B (8.12%) compared to the zero-shot setting. Mistral-7B, specifically tuned on our dataset, excels in precision for both "permit" and "forbid" cases, surpassing ChatGPT and approaching GPT-4's performance. However, using "Direct Prompt" result in a notable decline for Mistral-7B, from 66.98% to 51.75%, indicating limited grounding ability. Direct training on abstract legal concepts leads to reasoning confusion, as seen with Llama2-7B, which tends to misclassify cases as "forbid," (see Table 13). Our results reaffirm the high quality of cases generated under contextual integrity theory and the feasibility of the reasoning pipeline for adjudicating privacy law cases.

5.3 Ablation Study

To better understand how to ensure the quality of synthetic cases grounded in real law, we conduct several ablation studies. These studies demonstrate the effectiveness of our contextual feature filter, consistency checks, and diversity ranking. The complete results of these ablation studies are presented in Table 11.

Contextual Feature Filter We conduct ablation studies to assess the effect of contextual feature filters. After generating case backgrounds, we retain all cases including those that lacked key features (*e.g.*, *sender*, *recipient*) of contextual integrity. The results, denoted as (◇ w/o Feature F), reveal significant performance declines. Specifically, there is a drop of 3.26% and 2.36% in the applicability and compliance tasks, respectively, for Llama2-13B (see Table 3). These findings demonstrate the importance of feature integrity.

Consistency Filter First, we remove the norm consistency filter (◇ w/o Norm F) and do not verify whether the legal norms in synthetic cases match the seed norms. Here, Mistral-7B drops by 3.62% in the compliance task illustrating the efficacy of the norm consistency checker in mitigating issues

Models	Norm.$_{Acc}$		Conc.$_{Ma-F1}$	
	w/ CI	w/o CI	w/ CI	w/o CI
MPT-7B	34.58	29.91	58.62	53.44
Llama2-7B	46.73	39.25	59.58	56.72
Mistral-7B	51.40	45.79	66.98	61.22
Llama2-13B	53.27	43.93	64.83	59.69

Table 4: Performance comparison with and without contextual feature extraction in the first step during tuning and evaluation. **Norm.**$_{Acc}$ denotes norm retrieval accuracy, and **Conc.**$_{Ma-F1}$ indicates Macro F1-scores of conclusions (permit, forbid).

such as hallucinations during generation. Subsequently, we observe a significant performance decline when we bypass the check of the conclusion ($\diamond$ w/o Conclusion F). Incorrect conclusions lead to increased perplexity in legal judgments during training, which in turn causes a 4.99% drop in the applicability judgments for Llama2-13B.

Diversity Ranking We remove the diversity ranking ($\diamond$ w/o Diversity R) and randomly sample cases for each norm. Low diversity often results in high similarity among cases, such as in the roles of entities or specific categories of information. The lack of diversity can decrease the robustness of training, as demonstrated in (Wang et al., 2024, 2023). This impact is further reflected in a 3.86% decline in the Macro F1-score for applicability judgments in Llama2-13B. Furthermore, we deactivate all of the above filters and ranking mechanisms ($\diamond$ w/o All Parts) and observe significant decreases across all language models, with Mistral-7B experiencing drops of 5.89% and 5.05% in each task respectively. These findings underscore the importance of enhancing the integrity, consistency, and diversity of generated cases.

5.4 Discussion of GoldCoin Instruction

To further investigate whether the improvement in model performance stems from the quality of synthetic cases or the instructions themselves, we conduct experiments utilizing multi-step instructions on all baseline models (see results in Table 14). Additionally, we discuss how contextual integrity affects norm retrieval accuracy and judgment performance as shown in Table 4.

Multi-step Instruction As shown in Table 14, we can compare this with Table 2 and notice that the Macro F1-scores for MPT-7B and Mistral-7B exhibit a slight average improvement of 1.70% when determining the applicability of HIPAA under the zero-shot setting. Nonetheless, the Llama2 series shows a decline of 2.17%, indicating unstable performance when not aligned with specific cases. Similar results are reflected in the compliance tasks, demonstrating that merely relying on detailed instructions is insufficient to guide LLMs to follow contextual integrity for effective judgment. The instability may arise when the models are not exposed to such case types and legislation during pre-training, underscoring the importance of our approach that utilizes synthetic cases grounded in actual laws.

Features in Contextual Integrity Contextual Integrity (CI) (Nissenbaum, 2004) serves as a bridge between abstract privacy laws and specific cases, enhancing norm retrieval and subsequently improving judgment capabilities. We omit the contextual feature extraction step in the compliance task (w/o CI), and the results are presented in Table 4. The norm retrieval accuracy declines significantly across all open-source LLMs tuned by GoldCoin, demonstrating that contextual features effectively aid the model in understanding information transmission within cases and aligning them with pertinent legal statutes. Llama2-13B, which exhibits the best norm retrieval performance, experiences a significant decrease of 5.14% in conclusion performance when contextual integrity features are not extracted. These findings substantiate that contextual integrity is an effective formalization method in the privacy domain, further demonstrating the efficacy of our GoldCoin framework in aligning LLMs with privacy laws.

6 Conclusion

In this paper, we introduce GoldCoin, a pioneering framework that leverages the contextual integrity theory to effectively apply privacy laws to privacy violation detection. Specifically, we practice the HIPAA Privacy Rule and build synthetic cases for aligning LLMs. Our experimental results demonstrate that this approach significantly enhances models' capability to assess legal relevance and pinpoint privacy risks, providing a novel perspective for the integration of privacy legislation within LLMs. In the future, this generation and alignment method could be extended to other privacy laws such as GDPR and COPPA, or general legal domains. We hope our GoldCoin sheds light on the development of legal LLMs.

Limitations

Although our approach considers all `permit` and `forbid` norms within HIPAA, it does not account for interconnections between these norms. In practice, legal norms often contain cross-references and a single case may be adjudicated based on multiple norms (Lam et al., 2009). Future work should construct cases based on multiple norms to better reflect real-world scenarios and potentially yield improvements. Additionally, we do not consider the few-shot setting due to multiple examples often exceeding the maximum input length of LLMs. For the selection of laws, we conduct experiments on HIPAA due to its prominence in the privacy domain and the relatively abundant availability of open-source cases, which can serve as ground truth for testing. We invite legal professionals with access to cases related to other privacy laws to contact us, as this would facilitate the extension of our approach to additional privacy regulations such as COPPA (Aftab and Savitt, 1999), GDPR (Voigt and Von dem Bussche, 2017), etc. Moreover, this paper primarily focuses on case generation, and we do not employ techniques such as retrieval-augmented generation (Gao et al., 2024; Lewis et al., 2020) or vector embedding (Douze et al., 2024) for retrieving relevant norms. We believe that dynamically indexing (Liu, 2022) and retrieving related norms, based on the statute graph constructed in Section 3.1, is a promising direction.

Ethics Statement

We use the API provided by the Code of Federal Regulations official website to access the HIPAA Privacy Rule. We follow contextual integrity theory (Nissenbaum, 2004) to generate synthetic cases for constructing GOLDCOIN-HIPAA, and manually remove cases that could potentially leak real-world private information. We follow the official usage and access rules of the Caselaw Access Project[4] during downloading relevant cases. Human evaluations and annotations are performed by two legal experts to review the quality of synthetic cases and remove cases that contain explicit content such as gore or violence. Annotations are compensated at 15 USD per hour, above the local minimum wage. To the best of our knowledge, this work complies with open-source agreements and does not pose risks of information leakage or other hazards.

[4] https://case.law/about/#usage-access

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Accountability Act. 1996. Health insurance portability and accountability act of 1996. *Public law*, 104:191.

Parry Aftab and Nancy L Savitt. 1999. The children's online privacy protection act of 1998. *Preventive L. Rep.*, 18:32.

AI Anthropic. 2024. Introducing the next generation of claude.

Paul Ashley, Satoshi Hada, Günter Karjoth, Calvin Powers, and Matthias Schunter. 2003. Enterprise privacy authorization language (epal). *IBM Research*, 30:31.

Paul Ashley, Satoshi Hada, Günter Karjoth, and Matthias Schunter. 2002. E-p3p privacy policies and privacy authorization. In *Proceedings of the 2002 ACM workshop on Privacy in the Electronic Society*, pages 103–109.

Adam Barth, Anupam Datta, John C Mitchell, and Helen Nissenbaum. 2006. Privacy and contextual integrity: Framework and applications. In *2006 IEEE symposium on security and privacy (S&P'06)*, pages 15–29. IEEE.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *CoRR*, abs/2005.14165.

Felix Chang, Erin McCabe, and James Lee. 2020. Mining the harvard caselaw access project. *Available at SSRN 3529257*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law. *Preprint*, arXiv:2403.03883.

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.

Henry DeYoung, Deepak Garg, Limin Jia, Dilsun Kaynar, and Anupam Datta. 2010. Experiences in the logical specification of the hipaa and glba privacy laws. In *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society*, pages 73–82.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *Preprint*, arXiv:2401.08281.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, Heng Wang, and Zhiyuan Liu. 2019. *CJRC: A Reliable Human-Annotated Benchmark DataSet for Chinese Judicial Reading Comprehension*, page 439–451. Springer International Publishing.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022a. Unnatural instructions: Tuning language models with (almost) no human labor. *Preprint*, arXiv:2212.09689.

Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. 2022b. Instruction induction: From few examples to natural language task descriptions. In *Annual Meeting of the Association for Computational Linguistics*.

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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Richard Kuhn, Edward Coyne, and Timothy Weil. 2010. Adding attributes to role-based access control.

Peifung E Lam, John C Mitchell, and Sharada Sundaram. 2009. A formalization of hipaa for a medical messaging system. In *International Conference on Trust, Privacy and Security in Digital Business*, pages 73–85. Springer.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Conference on Empirical Methods in Natural Language Processing*.

Jerry Liu. 2022. LlamaIndex.

Kirsten Martin and Helen Nissenbaum. 2016. Measuring privacy: An empirical test using context to expose confounding variables. *Colum. Sci. & Tech. L. Rev.*, 18:176.

Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.

MosaicML NLP Team. 2023. Introducing mpt-30b: Raising the bar for open-source foundation models. Accessed: 2023-06-22.

Ha-Thanh Nguyen. 2023. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3. *Preprint*, arXiv:2302.05729.

Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–158.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Stuart L Pardau. 2018. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68.

Livio Robaldo, Cesare Bartolini, Monica Palmirani, Arianna Rossi, Michele Martoni, and Gabriele Lenzini. 2020. Formalizing gdpr provisions in reified i/o logic: the dapreco knowledge base. *Journal of Logic, Language and Information*, 29:401–449.

Ravi S Sandhu. 1998. Role-based access control. In *Advances in computers*, volume 46, pages 237–286. Elsevier.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chandan Singh, John X. Morris, Jyoti Aneja, Alexander M. Rush, and Jianfeng Gao. 2023. Explaining patterns in data with language models via interpretable autoprompting. *Preprint*, arXiv:2210.01848.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

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. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Y. Wong, and Simon See. 2024. Absinstruct: Eliciting abstraction ability from llms through explanation tuning with plausibility estimation. *Preprint*, arXiv:2402.10646.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *Preprint*, arXiv:2109.09193.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Seonghyeon Ye, Doyoung Kim, Joel Jang, Joongbo Shin, and Minjoon Seo. 2022. Guess the instruction! flipped learning makes language models stronger zero-shot learners. *ArXiv*, abs/2210.02969.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021a. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 973–982.

Linan Yue, Qi Liu, Han Wu, Yanqing An, Li Wang, Senchao Yuan, and Dayong Wu. 2021b. Circumstances enhanced criminal court view generation. SIGIR '21, page 1855–1859, New York, NY, USA. Association for Computing Machinery.

Han Zhang, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023. Contrastive learning for legal judgment prediction. *ACM Trans. Inf. Syst.*, 41(4).

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. *Preprint*, arXiv:2211.01910.

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 Contextual Integrity: Theory and Framework

In this appendix, we explore the concept of contextual integrity as developed by Nissenbaum (2004). This theory serves as a framework (Barth et al., 2006) for formalizing information transmission, particularly within various societal contexts.

A.1 Information Transmission

Information transmission involves three primary entities: the ***sender*** of the message, the ***recipient*** who receives the information, and the ***subject*** who is related to the information, also referred to as the ***about***. The ***information type*** $t \in \mathcal{T}$ is another crucial element, referring to the specific category of the transmitted information (*e.g.*, health plan, address). These elements constitute the fundamental components of transmission.

A.2 Roles and Contexts

At the core of contextual integrity lies the concept of **context**. Nissenbaum emphasizes that individuals operate not merely as undifferentiated entities but in specific roles within different social contexts, such as healthcare, education, employment, and marketplaces. Each entity within a context plays specific roles $r \in \mathcal{R}$. Understanding these roles is crucial as they significantly influence the nuanced judgments individuals make concerning potential privacy violations. For instance, Mr. Smith, depicted in Figure 1, may act as a doctor within a healthcare setting, subject to HIPAA (Act, 1996), a consumer in a supermarket, subject to the CCPA (Pardau, 2018), or a father within his family setting. Each role carries distinct expectations and norms regarding privacy. Accurately identifying and comprehending the role of entities within the specific context is essential for determining the appropriate law to apply in privacy risk detection.

A.3 Transmission Principles

After understanding the concept of information flows and context, we then expand to the concept of ***transmission principle***, which is a distinctive aspect of the contextual integrity approach to privacy. These principles define the specific constraints regulating the flow of information from one entity to another. In this work, we select ***Purpose, In Reply To, Consented By, Belief*** as the key transmission principles. The meanings of these principles are shown in Appendix C.1. Future extensions to other privacy legislations could involve adding new principles manually or guiding LLMs to automatically induce principles based on the target laws.

A.4 Informational Norms

With all features of contextual integrity in place, we introduce the concept of ***norm***. Norms governing the transmission of personal information from one party to another, referred to as "informational norms", are derived from societal expectations and legal standards. These norms restrict, for example, what physicians can disclose about the health conditions of patients under their care. Since societal expectations are challenging to define and subjective, this work relies on standardized legal frameworks to extract norms. We can represent a norm of information flow as $(\mathcal{P}, \mathcal{R}) \wedge \mathcal{T} \wedge \Omega$. Legal regulations such as HIPAA provide a formal definition for each type of information transmission, as expressed abstractly in Equation (1). Then the legality of each information transmission can be defined as:

$$\begin{aligned}&\text{inrole}(p_s, \hat{r}_s) \wedge \text{inrole}(p_r, \hat{r}_r) \wedge \text{inrole}(p_a, \hat{r}_a) \\ &\wedge (t \in \hat{t}) \wedge \Omega \rightarrow \{\text{permit, forbid}\},\end{aligned} \tag{4}$$

where p_s, p_r, p_a represent the sender, recipient, and subject, respectively. Besides, current research (Mireshghallah et al., 2023) explores expressing personal privacy expectations as norms to assess privacy risks. This approach also represents a valuable area for further exploration.

A.5 Example

To illustrate the application of norms, we consider the example from Figure 1. With the theory of contextual integrity, we map the features of the healthcare context to the formal representation in Equation (4) as follows:

$$\begin{aligned}&\text{inrole}(p_s, \text{doctor}) \wedge \text{inrole}(p_r, \text{doctor}) \wedge \\ &\text{inrole}(p_a, \text{patient}) \wedge (t \in \text{blood test results}) \wedge \\ &(\omega_{purp} \in \text{treatment planning}),\end{aligned} \tag{5}$$

where ω_{purp} denotes the purpose of the information transmission. Given that the doctor is a covered entity under HIPAA and blood test results are health information, this information transmission aligns with the legal norm `164.502(a)(1)(ii)` of HIPAA, thereby being permitted.

B HIPAA Privacy Rule

B.1 Brief Introduction

The Health Insurance Portability and Accountability Act (HIPAA) of 1996 Title II sets national standards for protecting personal health information (PHI). Defined as PHI, this includes any individually identifiable health information managed by entities such as health plans, health care clearinghouses, and health care providers who transmit health information electronically. The HIPAA Privacy Rule, detailed in `45 CFR Parts 160, 162, and 164`, provides federal protections for PHI, limits its disclosure without consent, and gives patients rights regarding their health information, such as accessing and amending their records. The Privacy Rule is codified at `45 CFR Part 160`[5] and `Subparts A and E of Part 164`[6].

B.2 Requirement, Exception, and General Definition

In Section 3.1, besides norms like **`permit`** and **`forbid`** that describe compliance with privacy information transmission, HIPAA also includes the following basic types of norms:

- **`Requirement`** indicates that an action is permissible under the rule only if specific conditions are met. For example, according to `164.508(a)(2)`, an action is allowable only with proper authorization.
- **`Exception`** refers to a specific scenario where a standard rule or requirement does not need to be applied. For instance, `164.508(a)(2)` specifies that if psychotherapy notes are used for the treatment of the originator, the usual authorization requirement is waived.
- **`General definition`** provides a broad explanation of concepts or terms. For example, in HIPAA terminology, a "covered entity" is defined as a health plan, a health care clearinghouse, or a health care provider who transmits health information in electronic form.

Single norm may consist of multiple types (e.g., permit with requirement, permit with exception). In this work, we focus only on norms containing the permit and forbid types.

[5] https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-160?toc=1

[6] https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164

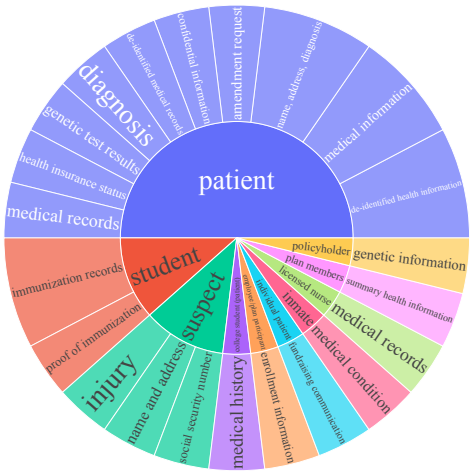

Figure 7: Top 10 common information subjects (inner circle) and their corresponding top 10 information types (outer circle).

♦ Applicability	# Train	# Test
Synthetic (Applicable)	309	-
Synthetic (Not Applicable)	-	-
Real (Applicable)	-	107
Real (Not Applicable)	309	107
♦ Compliance	**# Train**	**# Test**
Synthetic (Permit)	269	-
Synthetic (Forbid)	40	-
Real (Permit)	-	80
Real (Forbid)	-	27

Table 5: Statistics of GOLDCOIN-HIPAA.

B.3 Details of Norm Classification

In this appendix, we provide the prompt for norm classification in Table 15. We compile each norm with the classification instruction and utilize GPT-4 to extract the basic norm type.

Statistics In this work, we mainly focus on the `45 CFR Part 164`, which governs security and privacy concerns in the healthcare sector. Following the classification in Section 3.1, we analyze the types within each norm, the statistical results are presented in Table 6. We have identified 269 of 691 HIPAA norms that permit certain information transmission and 40 of 691 norms that prohibit specific transmission.

C Details of GOLDCOIN-HIPAA Dataset

C.1 Prompt of Background Generation

To ensure that contextual integrity is considered when constructing case backgrounds, we incorporate the definition of privacy information flow along

Norm Type	# Number
Total	691
- Permit	**269**
- Forbid	**40**
- Requirement	555
- Exception	112
- Definition	44

Table 6: Statistics of norm categories within HIPAA Privacy Rule. Each norm may encompass several norm types, thereby the **Total** number of norms is less than the cumulative sum of individual types.

with key contextual features into the prompt, as shown in Table 16. Based on the provided norm (i.e., regulation, clause) and its type (e.g., `permit`, `forbid`), we prompt GPT-4 to generate the corresponding case background.

C.2 Statistics

This appendix presents the statistics of the training and testing datasets used in this study. The term "Synthetic" refers to cases generated by GOLDCOIN, which are based on HIPAA regulations, while "Rea" indicates cases collected from CAP (see Appendix D) and processed through our pipeline. The statistics of GOLDCOIN-HIPAA dataset are provided in Table 5.

C.3 Case Study

We present two examples to visually show the quality of the cases generated by our framework. The first example is a case permitted under `164.502(j)(1)(i)`, as shown in Figure 8. The second example is a case forbidden by the `164.502(a)(5)(ii)(B)(1)`, as detailed in Figure 9. Each case includes one seed norm, a background story, features related to contextual integrity, and a conclusion.

D Details of the Caselaw Access Project

The Caselaw Access Project (CAP), an initiative by the Harvard Law School Library, has digitized a comprehensive collection of American case law. This monumental effort has converted approximately 40 million pages of court decisions into a machine-readable format, thus making these legal documents accessible online in a consistent format. The collection includes all official, book-published state and federal U.S. case law up to the year 2020, covering a wide range of courts, including state, federal, and territorial courts.

D.1 Dataset Collection

We utilized the official API[7] provided by CAP, employing "HIPAA Privacy Rule" as the keyword for dumping relevant cases. We filtered out cases longer than 30,000 words and shorter than 100 words before proceeding with further processing. Additionally, we sampled 2,000 cases related to general privacy violations using the keyword "privacy violation" to provide a training and testing set for the applicability task.

D.2 Prompt

In this appendix, we provide the prompt as depicted in Table 17 for case processing by GPT-4. Since a real case may relate to other legal regulations except HIPAA, we target to extract the factual background, contextual features, related norms, and court conclusions that are relevant to the HIPAA Privacy Rule.

D.3 Human Annotation

After the preliminary processing by GPT-4, we engaged two human experts who had studied privacy protection and privacy laws for over a year to manually annotate, correct, and filter the HIPAA-related cases. The annotations focused on three main tasks:(1) Removing cases not related to information transmission. (2) Deleting the court analysis from the background. (3) Assessing whether the court conclusions were correctly extracted.

D.4 Case Study

In this appendix, we present three real court cases processed by our pipeline. The first case is an example where HIPAA permits the transmission, as shown in Figure 11. The second case is an example where HIPAA forbids the transmission, as detailed in Figure 12. The third case is an example that HIPAA is not applicable, as outlined in Figure 10.

E Implement Details

We select four open-source language models that support at least 4K tokens input and instruction-tune them on one H800 (80G) GPU. Specifically, we parameter-efficient fine-tune MPT-7B, Mistral-7B, Llama2-7B, and Llama2-13B using LoRA. For LoRA, we choose a rank and alpha of 8 and 16, respectively. All language models are trained for 3 epochs, and we select the final checkpoints for evaluation. The batch size is 1, and the learning

[7] https://old.case.law/docs/site_features/api

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
{instruction}

Input:
{input}

Response:

(a) Template for examples with an input.

Below is an instruction that describes a task. Write a response that appropriately completes the request.

Instruction:
{instruction}

Response:

(b) Template for examples with an empty input.

Table 7: The prompt templates used to concatenate instructions and example inputs. Two templates are shown to account for cases where the input is optional. Placeholders **instruction** and **input** are replaced with actual instructions and inputs.

rate is set to 1e-5. For API-based LLMs, we access ChatGPT and GPT-4 via the Azure OpenAI API Service[8], using the versions `gpt-3.5-turbo (2024-02-01)` and `gpt-4 (2024-02-01)`. The total generation and evaluation costs of using the API are approximately $100 and $20. respectively.

E.1 Instruction Template

To align the knowledge in our case instructions with the language models without compromising its overall performance, we follow the approach described in (Taori et al., 2023). The specific prompt used as the instruction template can be found in Table 7.

E.2 Vanilla and Multi-step Prompts for the Applicability Task

This appendix details the prompts utilized in the applicability task. For the "Direct Prompt" along with other baseline approaches, we employed the vanilla prompt, as illustrated in Table 8(a). The response format for the vanilla prompt is straightforward, consisting of either "Applicable" or "Not Applicable". Additionally, we implemented multi-step prompts for GOLDCOIN, which are depicted in Table 8(b).

[8] https://azure.microsoft.com/en-us/products/ai-services/openai-service

Instruction: Please determine whether the HIPAA Privacy Rule is applicable to the case.

Input: Read the case background: <background>.

Response: Applicable / Not applicable.

(a) Vanilla prompt of the applicability task.

Instruction: Please assess the applicability of the HIPAA Privacy Rule to the case through the following steps: Step 1: Annotate the message characteristics [Sender, Sender Role, Recipient, Recipient Role, Subject, Subject Role, Type] about the flow of private information in the case as a list. Step 2: Determine whether the HIPAA Privacy Rule is applicable to the case.

Input: Read the case background: <background>.

Response:
Step 1: Sender: <sender>, Sender Role: <sender role> ...
Step 2: Applicable / Not applicable.

(b) Multi-step instruction of the applicability task.

Table 8: The vanilla and multi-step prompts we used in the applicability task.

E.3 Vanilla and Multi-step Prompts for the Compliance Task

In the compliance task, we utilize the same prompt settings as those in the applicability task. For "Direct Prompt", the response to the vanilla prompt is limited to either "Permit" or "Forbid". This setting is depicted in Table 9(a). For the multi-step instruction presented in Table 9(b), we instruct language models to analyze contextual features and retrieve relevant norms for making informed judgments.

E.4 Prompt of Law Recitation

To align language models with the content from privacy laws as a baseline, we build an instruction that guides the models to recite the content of the HIPAA Privacy Rule. As shown in Table 10, we incorporate all norms from HIPAA `Part 164` into this template for tuning.

F Supplementary Experiments

This section provides additional supplementary experiments to Section 5.

F.1 Full Results of Ablation Study

In this appendix, we provide an overall comparison of four LLMs under different ablation settings, focusing on their specific performance deficits as detailed in Table 11. It is observed that inaccuracies in conclusion lead to the most substantial performance degradation, particularly for MPT-7B, which experiences a 6.25% reduction in accuracy

Instruction: Please determine whether the HIPAA Privacy Rule permits or forbids the case.
Input: Read the case background: <background>.
Response: Permit / Forbid.

(a) Vanilla prompt of the compliance task.

Instruction: Please assess the case for compliance with the HIPAA Privacy Rule through the following steps: Step 1: Annotate the eleven message characteristics [Sender, Sender Role, Recipient, Recipient Role, Subject, Subject Role, Type, Purpose, In Reply To, Consented By, Belief] about the flow of private information in the case as a list. Step 2: Identify and list all applicable HIPAA regulation IDs (e.g., 164.xxx) and their content. Step 3: Determine whether the HIPAA Privacy Rule permits or forbids the case.
Input: Read the case background: <background>
Response: Step 1: Sender: <sender>, Sender Role: <sender role> ... Step 2: <norm id>, <norm content>. Step 3: Permit / Forbid.

(b) Multi-step instruction of the compliance task.

Table 9: The vanilla and multi-step prompts we used in the compliance task.

Instruction: Please recite the contents of <norm id> in the HIPAA Privacy Rule.
Response: <norm content>.

Table 10: Prompt of the baseline "Law Recitation".

Model	Applicability	Δ_{App}	Compliance	Δ_{Com}
MPT-7B	**67.30**	-	**58.62**	-
◇ w/o Feature F	65.39	1.91↓	57.87	0.75↓
◇ w/o Norm F	66.28	1.02↓	55.43	3.19↓
◇ w/o Conclusion F	61.05	6.25↓	53.75	4.87↓
◇ w/o Diversity R	64.28	3.02↓	56.27	2.35↓
◇ w/o All Parts	60.48	6.82↓	53.14	5.48↓
Llama2-7B	**94.39**	-	**59.58**	-
◇ w/o Feature F	92.88	1.51↓	57.94	1.64↓
◇ w/o Norm F	93.74	0.65↓	56.23	3.35↓
◇ w/o Conclusion F	91.03	3.36↓	54.56	5.02↓
◇ w/o Diversity R	92.15	2.24↓	56.85	2.73↓
◇ w/o All Parts	89.06	5.33↓	53.02	6.56↓
Mistral-7B	**97.66**	-	**66.98**	-
◇ w/o Feature F	95.22	↓2.44	65.04	↓1.92
◇ w/o Norm F	95.98	↓1.68	63.34	↓3.62
◇ w/o Conclusion F	93.61	↓4.05	63.05	↓3.91
◇ w/o Diversity R	95.54	↓2.12	64.45	↓2.51
◇ w/o All Parts	91.77	↓5.89	61.91	↓5.05
Llama2-13B	**99.53**	-	**64.83**	-
◇ w/o Feature F	96.27	↓3.26	62.47	↓2.36
◇ w/o Norm F	97.59	↓1.94	61.34	↓3.49
◇ w/o Conclusion F	94.54	↓4.99	61.07	↓3.76
◇ w/o Diversity R	95.67	↓3.86	62.33	↓2.50
◇ w/o All Parts	93.01	↓6.52	60.11	↓4.72

Table 11: Ablation study for MPT-7B, Llama2-7B, Mistral-7B and Llama2-13B. Macro F1-scores are exhibited, and Δ_{All} indicates score changes.

when determining applicability. The performance loss due to inconsistencies in norms reveals that GPT-4 continues to manifest certain hallucinatory and random behaviors during case generation.

F.2 Overall Performance across Categories

In this appendix, we extend the experimental results introduced in Table 2 across four LLMs and GPT series. Table 12 provides a comprehensive dataset of experimental results for the applicability task using "Zero-shot," "Law Recitation," "Direct Prompt," and our proposed method GOLDCOIN. Metrics such as Precision (Prec), Recall (Rec), and F1-score (F1) are evaluated for both "Applicable" and "Not Applicable" categories, alongside average Accuracy (Acc) and Macro F1-score (Ma-F1). GPT-4 exhibits optimal performance with the "Direct Prompt", whereas its efficacy declines when employing multi-step instructions, corroborating the findings discussed in Section 5.4. Utilizing GOLDCOIN, Mistral-7B and Llama2-13B achieve 100% in precision for the positive category and recall in the negative category. We also provide a detailed analysis of the compliance task as shown in Table 13 and the inherent instability of the "Direct Prompt" is evident; for instance, Mistral-7B reached a precision of 97.44% in the "permit" category, yet the precision for "forbid" was merely 27.94%. These findings underscore the necessity of integrating our multi-step instructions with the generated cases to achieve optimal outcomes.

F.3 Baselines under Multi-step Instruction

Table 14 outlines the performances when multi-step instructions are integrated into all baseline models. As discussed in Section 5.4, the direct application of multi-step prompting in LLMs without instruction-tuning on GOLDCOIN-HIPAA results in performance degradation. Notably, Llama-2 13B exhibits a 3.33% decrease in the "Zero-shot" setting. This decline is attributed to the model's inability to comprehend and apply contextual integrity without direct reference to legal knowledge. Furthermore, the top sections of Table 13 and Table 12 illustrate how GPT models fare when subjected to multi-step instruction scenarios.

Method	Models	Applicable			Not Applicable			All	
		Prec	Rec	F1	Prec	Rec	F1	Acc	Ma-F1
LLM API	ChatGPT	94.90	86.92	90.73	87.93	95.33	91.48	91.12	91.11
	GPT-4	97.17	96.26	96.71	96.30	97.20	96.74	**96.73**	**96.73**
	ChatGPT (MS)	95.00	88.79	91.79	89.47	95.33	92.31	92.06	92.05
	GPT-4 (MS)	92.79	96.26	94.50	96.12	92.52	94.29	94.39	94.39
Zero-shot	MPT-7B	55.08	60.75	57.78	56.25	50.47	53.20	55.61	55.49
	Llama2-7B	65.22	98.13	78.36	96.23	47.66	63.75	72.90	71.05
	Mistral-7B	91.18	86.92	89.00	87.50	91.59	89.50	89.25	89.25
	Llama2-13B	98.89	83.18	90.36	85.48	99.07	91.77	**91.12**	**91.07**
Law Recitation	MPT-7B	44.21	39.25	41.58	45.38	50.47	47.79	44.86	44.69
	Llama2-7B	66.46	98.13	79.25	96.43	50.47	66.26	74.30	72.75
	Mistral-7B	88.89	82.24	85.44	83.48	89.72	86.49	85.98	85.96
	Llama2-13B	95.88	86.92	91.18	88.03	96.26	91.96	**91.59**	**91.57**
Direct Prompt	MPT-7B	100.00	27.10	42.65	57.84	100.00	73.29	63.55	57.97
	Llama2-7B	100.00	78.50	87.96	82.31	100.00	90.30	89.25	89.13
	Mistral-7B	100.00	90.65	95.10	91.45	100.00	95.54	**95.33**	**95.32**
	Llama2-13B	97.03	91.59	94.23	92.04	97.20	94.55	94.39	94.39
GOLDCOIN	MPT-7B	77.46	51.40	61.80	63.64	85.05	72.80	68.22	67.30
	Llama2-7B	97.03	91.59	94.23	92.04	97.20	94.55	94.39	94.39
	Mistral-7B	100.00	95.33	97.61	95.54	100.00	97.72	97.66	97.66
	Llama2-13B	100.00	99.07	99.53	99.07	100.00	99.53	**99.53**	**99.53**

Table 12: Performance of GOLDCOIN and baselines under different settings across "Applicable" and "Not Applicable" categories. We **bold** the best results and underline the second-best results in each setting. **MS** denotes the setting of employing multi-step instruction.

Method	Models	Permit			Forbid			All	
		Prec	Rec	F1	Prec	Rec	F1	Acc	Ma-F1
LLM API	ChatGPT	88.00	75.86	81.48	34.38	55.00	42.31	71.96	61.89
	GPT-4	87.21	86.21	86.71	42.86	45.00	43.90	**78.50**	65.30
	ChatGPT (MS)	86.59	81.61	84.02	36.00	45.00	40.00	74.77	62.01
	GPT-4 (MS)	92.86	74.71	82.80	40.54	75.00	52.63	74.77	**67.72**
Zero-shot	MPT-7B	77.78	48.28	59.57	15.09	40.00	21.92	46.73	40.75
	Llama2-7B	81.25	59.77	68.87	18.60	40.00	25.40	56.07	47.14
	Mistral-7B	94.74	41.38	57.60	26.09	90.00	40.45	50.47	49.02
	Llama2-13B	86.76	67.82	76.13	28.21	55.00	37.29	**65.42**	**56.71**
Law Recitation	MPT-7B	70.37	43.68	53.90	7.55	20.00	10.96	39.25	32.43
	Llama2-7B	86.11	35.63	50.41	21.13	75.00	32.97	42.99	41.69
	Mistral-7B	78.46	58.62	67.11	14.29	30.00	19.35	53.27	43.23
	Llama2-13B	88.41	70.11	78.21	31.58	60.00	41.38	**68.22**	**59.79**
Direct Prompt	MPT-7B	85.92	70.11	77.22	27.78	50.00	35.71	66.36	56.46
	Llama2-7B	85.07	65.52	74.03	25.00	50.00	33.33	62.62	53.68
	Mistral-7B	97.44	43.68	60.32	27.94	95.00	43.18	53.27	51.75
	Llama2-13B	87.34	79.31	83.13	35.71	50.00	41.67	**73.83**	**62.40**
GOLDCOIN	MPT-7B	86.49	73.56	79.50	30.30	50.00	37.74	69.16	58.62
	Llama2-7B	84.21	91.95	87.91	41.67	25.00	31.25	**79.44**	59.58
	Mistral-7B	90.67	78.16	83.95	40.62	65.00	50.00	75.70	**66.98**
	Llama2-13B	87.80	82.76	85.21	40.00	50.00	44.44	76.64	64.83

Table 13: Performance of GOLDCOIN and baselines under different settings across "Permit" and "Forbid" categories. We **bold** the best results and underline the second-best results in each setting. **MS** denotes the multi-step instruction.

Task	Method	MPT-7B		Llama2-7B		Mistral-7B		Llama2-13B	
		Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1
Applicability	Zero-shot	57.01	57.01	70.09	70.03	91.12	91.11	87.85	87.74
	Law Recitation	44.86	44.82	71.03	70.82	89.72	89.70	92.06	92.05
	GOLDCOIN	68.22	67.30	94.39	94.39	<u>97.66</u>	<u>97.66</u>	**99.53**	**99.53**
Compliance	Zero-shot	48.60	41.80	57.01	46.73	57.01	49.61	67.29	54.95
	Law Recitation	42.99	37.39	48.60	41.18	53.27	45.23	67.29	58.15
	GOLDCOIN	69.16	58.62	**79.44**	59.58	75.70	**66.98**	<u>76.64</u>	<u>64.83</u>

Table 14: Performance of four LLMs with multi-step instruction setting, showing **Acc** and **Ma-F1** across both applicability and compliance tasks. We **bold** the best results and <u>underline</u> the second-best results in each task.

Now you are a legal expert on HIPAA Privacy Rule that answers questions as simply as possible.
Please read the following regulation text, and finish the following task.

Q1: (Classification) Classify the regulation type of the following regulation. The regulation type is one of the following: "Definition", "Permit", "Forbid", "Exception", "Requirement", "Permit and Exception", "Forbid and Exception", "Permit and Requirement", "Forbid and Requirement", "Permit and Exception and Requirement", "Forbid and Exception and Requirement", "Other".
Definition: The regulation defines a term or characteristic.
Permit: The regulation permits certain actions regarding the flow of private information.
Forbid: The regulation forbids certain actions regarding the flow of private information.
Exception: The regulation defines an exception to a certain action about privacy information flow.
Requirement: The regulation defines a requirement for privacy information flow.
Other: The regulation is not in the above types.

Q2: If the regulation type is "Definition", please annotate the name of the term or characteristic defined in the regulation.

Q3: If the regulation type is "Definition", please annotate the definition of the term or characteristic defined in the regulation.

Q4: If the regulation type contains "Permit", please annotate the action permitted in the regulation.

Q5: If the regulation type contains "Forbid", please annotate the action forbidden in the regulation.

Q6: If the regulation type contains "Exception", please annotate the exception defined in the regulation.

Q7: If the regulation type contains "Requirement", please annotate the requirement defined in the regulation.

Q8: If the regulation type is "Other", please give your own classification of the regulation type.

Table 15: Prompt of classifying norm types (*i.e.*, categories). GPT-4 is further instructed to provide details of each category.

Now you are a legal expert on HIPAA Privacy Rule that answers questions as simply as possible.
Read the regulation: {text}.
The regulation type is: {type}.

Q1. Create a legal case only related to regulation id and the type type. The case must be a detailed story in plain text, spanning between 200 to 500 words, closely related to the regulation id. The story must include the following seven characteristics about the flow of private information: [Sender, Sender Role, Recipient, Recipient Role, Subject, Subject Role, Type].
The "Sender" and "Recipient" fields indicate the sender and recipient of the message. The "Sender Role" and "Recipient Role" fields indicate the role of the sender and recipient (e.g., doctor, patient). The "Subject" and "Subject Role" field identifies the subject whose personal health information is contained in the message and the role of the subject. The "Type" field defines what kind of information would be passed, such as name or location. Integrate these seven characteristics seamlessly into the story without explicitly listing them.
Except for the seven characteristics, you can add the four optional characteristics [Purpose, In Reply To, Consented By, Belief] if mentioned in the regulation id or necessary. The "Purpose" field indicates a reason the message is being sent, such as for medical treatment. The "In Reply To" field was added to describe a disclosure where the message is sent as a response to some earlier message. The "Consented By" field indicates which people have consented to the message disclosure. The "Belief" field contains a collection of assertions about the current situation, such as whether this is a medical emergency, or whether the disclosure is (in the opinion of the sender) in the best interest of the health of the patient. Integrate these four characteristics seamlessly into the story without explicitly listing them.

Q2: Based on the background created in Q3, list the eleven characteristics regarding the flow of private information (Mark as "None" if not exist)

Q3: Please retrieve all the specific HIPAA regulation IDs that are the permission or prohibition description of the case. Please be as specific as possible to the sub-section id (e.g., 164.xxx).

Q4: Please classify the relation between the case and the regulation id as one of the following: "Permit", "Forbid", "Not Applicable".

Q5: Please classify the relation between the case and the HIPAA Privacy Rule as one of the following: "Permit", "Forbid", "Not Applicable".

Table 16: Prompt of case generation. We guide GPT-4 to generate case backgrounds and other details through a series of questions.

§ Seed Norm Id: 164.502(j)(1)(i)
§ Seed Norm Content:
HIPAA: HIPAA Privacy Rule
Part164: PART 164 — SECURITY AND PRIVACY
Part164SubpartE: Subpart E—Privacy of Individually Identifiable Health Information
164.502:§ 164.502 Uses and disclosures of protected health information: General rules.
164.502(j):(j) Standard: Disclosures by whistleblowers and workforce member crime victims
164.502(j)(1):(1) Disclosures by whistleblowers. A covered entity is not considered to have violated the requirements of this subpart if a member of its workforce or a business associate discloses protected health information, provided that:
164.502(j)(1)(i):(i) The workforce member or business associate believes in good faith that the covered entity has engaged in conduct that is unlawful or otherwise violates professional or clinical standards, or that the care, services, or conditions provided by the covered entity potentially endangers one or more patients, workers, or the public.
§ Case Background: Jane, a nurse at Happy Valley Hospital, recently became aware of a situation where the hospital management was not following proper sterilization procedures for surgical instruments. She believed this could potentially endanger the health and safety of patients, other workers, and the public. Concerned about the situation, Jane decided to report the issue to the state health department. She gathered evidence, including protected health information about several affected patients, and sent it to a health department official, Tom. Jane believed in good faith that disclosing this information was necessary to report the hospital's misconduct. Tom, in his role at the health department, reviewed the information and initiated an investigation into the hospital's practices. In this case, the Sender is Jane, the Sender Role is a nurse, the Recipient is Tom, the Recipient Role is a health department official, and the Subject Role is patient. The Subject field includes the patients whose protected health information was shared in the disclosure. The Type of information shared includes protected health information related to the improper sterilization issue. The Purpose of the disclosure is to report the hospital's misconduct for violating professional and clinical standards.
§ Contextual Features:
Sender: Jane, Sender Role: Nurse, Recipient: Tom, Recipient Role: Health department official,
Subject: Affected patients, Subject Role: Patients, Type: Protected health information,
Purpose: Report hospital's misconduct, Belief: Disclosure is necessary to report hospital's misconduct
§ Compliance Conclusion: Permit

Figure 8: A synthetic case generated by GOLDCOIN complies with HIPAA Privacy Rule.

§ Seed Norm Id: 164.502(a)(5)(ii)(B)(1)
§ Seed Norm Content:
HIPAA: HIPAA Privacy Rule
Part164: PART 164 — SECURITY AND PRIVACY
Part 164 SubpartE: Subpart E—Privacy of Individually Identifiable Health Information
164.502:§ 164.502 Uses and disclosures of protected health information: General rules.
164.502(a):(a) Standard. A covered entity or business associate may not use or disclose protected health information, except as permitted or required by this subpart or by subpart C of part 160 of this subchapter.
164.502(a)(5):(5) Prohibited uses and disclosures
164.502(a)(5)(ii):(ii) Sale of protected health information:
164.502(a)(5)(ii)(B):(B) For purposes of this paragraph, sale of protected health information means:
164.502(a)(5)(ii)(B)(1):(1) Except as provided in paragraph (a)(5)(ii)(B)(2) of this section, disclosure of protected health information by a covered entity or business associate, if applicable, where the covered entity or business associate directly or indirectly receives remuneration from or on behalf of the recipient of the protected health information in exchange for the protected health information.
§ Case Background: Jane Doe, a patient at XYZ Clinic, recently underwent a medical procedure. Dr. Smith, her treating physician at the clinic, documented her medical history, diagnosis, and treatment plan in her electronic health records. XYZ Clinic has a partnership with a pharmaceutical company, PharmaCorp, which is interested in obtaining patient data for research purposes. The clinic's administrator, without Jane's knowledge or consent, sends her protected health information (PHI) to PharmaCorp in exchange for financial remuneration.PharmaCorp's researcher, who receives Jane's PHI, analyzes it to develop new drugs and treatment plans. The researcher is aware that the information has been obtained in exchange for payment to the clinic. Meanwhile, Jane learns about this transaction and is upset that her PHI has been shared without her consent. She files a complaint with the Department of Health and Human Services (HHS).
§ Contextual Feature: Sender: XYZ Clinic's administrator, Sender Role: Covered Entity, Recipient: PharmaCorp's researcher, Recipient Role: Business Associate, Subject: Jane Doe, Subject Role: Patient, Type: Protected Health Information (PHI), Purpose: Research
§ Compliance Conclusion: Forbid

Figure 9: A synthetic case generated by GOLDCOIN does not comply with HIPAA Privacy Rule.

§ Related Norm Id: None
§ Case Background: On July 10, 2006, the plaintiff filed a complaint against the United States pursuant to 26 U.S.C. § 7433 claiming that the Internal Revenue Service ("IRS") wrongfully disclosed her tax return information to the public. Compl., Miller v. United States, No. 06-cv-01250 (D.D.C.), 6-12. On August 28, 2006, the plaintiff filed a second suit against the United States based on the same alleged misconduct, only this time complaining that the defendant had violated 26 U.S.C. § 7431. Compl. at ¶ 1. The § 7433 and § 7431 actions proceeded in parallel until the former was dismissed for failure to state a claim on July 19, 2007. In her remaining § 7431 suits, the plaintiff seeks damages for "substantial personal embarrassment, loss of goodwill, loss in credit .... and actual damages totaling $65,000." Am. Compl. at 19. The court permitted an amendment to the plaintiffs complaint on September 18, 2006, because no responsive pleadings had yet been filed. See Fed.R.CivP. 15(a). On November 6, 2006, the defendant moved to dismiss the plaintiffs amended complaint for lack of subject-matter jurisdiction and for failure to state a claim. The plaintiff did not file a response, but on January 22, 2007, she filed a motion to amend her complaint again. The court now turns to the merits of the government's motion to dismiss the plaintiffs § 7431 claim and the plaintiffs motion to amend her complaint for a second time.
§ Contextual Feature: Sender: Internal Revenue Service (IRS), Sender Role: Government agency, Recipient: Public, Recipient Role: Public, Subject: Plaintiff's tax return information, Subject Role: Plaintiff, Type: Tax return information
§ Applicability Conclusion: Not Applicable

Figure 10: A real court case sourced from CAP and is not relevant to HIPAA.

Now you are a legal expert on HIPAA Privacy Rule that answers questions as simply as possible.
Read the case: {case}.

Q1. If the case involves the flow of private information. Please annotate the eleven message characteristics [Sender, Sender Role, Recipient, Recipient Role, Subject, Subject Role, Type, Purpose, In Reply To, Consented By, Belief] about the flow of private information in the case as a list. If the characteristic does not exist, just fill in None.
The "Sender" and "Recipient" fields indicate the sender and recipient of the message. The "Sender Role" and "Recipient Role" fields indicate the role of the sender and recipient (*e.g.*, doctor, patient). The "Subject" and "Subject Role" field identifies the subject whose personal health information is contained in the message and the role of the subject. The "Type" field defines what kind of information would be passed, such as name or location. The "Purpose" field indicates a reason the message is being sent, such as for medical treatment. The "In Reply To" field was added to describe a disclosure where the message is sent as a response to some earlier message. The "Consented By" field indicates which people have consented to the message disclosure. The "Belief" field contains a collection of assertions about the current situation, such as whether this is a medical emergency, or whether the disclosure is (in the opinion of the sender) in the best interest of the health of the patient.

Q2: Please retrieve all the specific HIPAA regulation IDs that are the permission or prohibition description of the case. Please be as specific as possible to the sub-section id (e.g., 164.xxx). If the regulations do not exist, just fill in None.

Q3: Please classify the type of regulation(s). The regulation type is one of the following: "Definition", "Permit", "Forbid", "Exception", "Requirement", "Permit and Exception", "Forbid and Exception", "Permit and Requirement", "Forbid and Requirement", "Permit and Exception and Requirement", "Forbid and Exception and Requirement", "Other".

Q4: Please classify the relation between the case and each regulation in Q3 as one of the following: "Permit", "Forbid", and "Not Applicable".

Q5: A case may be associated with multiple regulations. If it is permitted by some regulations and not forbidden by any of the regulations, the case complies with HIPAA, answer "Permit". If it is not permitted by any of the regulations or forbidden by some regulations, the case violates HIPAA, answer "Forbid". Otherwise, if the case is not applicable to HIPAA, answer "Not Applicable". Please classify the relation between the flow of private information in the case and HIPAA as one of the following: "Permit", "Forbid", and "Not Applicable".

Q6: With the eleven characteristics in Q2, restore the BACKGROUND story of the case, especially about the flow of private information.

The case should not include any information about the regulation(s) in Q2 and the court decision. Make sure that the eleven characteristics are obviously included in the BACKGROUND story. The background must be a detailed story in plain text, spanning between 200 to 500 words.

Table 17: Prompt of parsing real court cases sourced from CAP. We guide GPT-4 through multiple questions to automatically extract HIPAA-related background stories for subsequent manual annotation.

§ Related Norm Id: 164.512(c)(1)
§ Case Background: In Chester County, an 88-year-old woman, A.M., found herself at the center of a complex legal and medical situation. The county's Department of Aging Services received a confidential report suggesting that A.M., living alone, might be a victim of financial mismanagement or abuse. To investigate, the Department assigned Douglas Bernard as the investigator. Bernard, an experienced investigator, understood the gravity and sensitivity of his task. His role was to gather information on A.M.'s physical and mental health, as well as her financial situation, to determine if protective services were necessary. This included evaluating whether A.M. was capable of managing her affairs, especially in light of a diagnosis of dementia and hypertension. Initially, A.M. seemed cooperative. She agreed to meet her primary care physician, Dr. Priem and even gave consent for the release of her medical information. However, this cooperation was short-lived. Within days, she revoked the medical release and started to avoid any further engagement with the investigator. Bernard's attempts to visit her at home were met with silence; A.M. refused to answer her door. Despite these challenges, Bernard persisted in his investigation. He reached out to A.M.'s attorney, William McLaughlin, but found no help there either. A.M. had refused to grant McLaughlin consent to disclose any information to the investigator. In his report to the court, Bernard detailed his efforts and A.M.'s increasing uncooperativeness. He shared his concerns about A.M.'s mental health, citing her confusion and memory loss. He also expressed worry about her financial situation, noting the presence of large amounts of cash at her home and irregularities in her financial accounts. The court, faced with this information, had to make a critical decision. Was A.M. indeed at risk of abuse, neglect, or exploitation? Was she mentally and physically capable of managing her own affairs? And crucially, was the flow of her private information, without her explicit consent, justified in this context? A.M.'s case was a delicate balance between protecting her rights and ensuring her well-being.
§ Contextual Feature:
Sender: Douglas Bernard, Sender Role: Investigator, Recipient: Court, Recipient Role: Legal Authority, Subject: A.M. (the older adult), Subject Role: Subject of the Protective Services Investigation, Type: Medical and financial information, Purpose: To assess the need for protective services and to support the petition for involuntary intervention, In Reply To: The Department's requirement to assess A.M.'s situation under the Older Adults Protective Services Act, Consented By: None (A.M. revoked consent for the release of her medical information), Belief: The belief that A.M. is at imminent risk of abuse, neglect, exploitation, or abandonment and is incapacitated due to dementia
§ Compliance Conclusion: Permit

Figure 11: A real court case sourced from CAP complies with HIPAA Privacy Rule.

§ Related Norm Id: 164.512(e)(1)
§ Case Background: In December 1997, Richard Moss was involved in a traffic accident when Jennifer Amira rear-ended his vehicle. Moss, suffering from injuries, was taken to Northwest Community Hospital for immediate medical attention. Here, he received an examination, was fitted with a collar, given a prescription, and later released. Months later, in June 1998, Moss consulted Dr. Richard Moser, a neurological surgeon, for further evaluation. The encounter between Moss and Dr. Moser formed the basis of a subsequent medical opinion about the nature and cause of Moss's injuries. As the legal case progressed, defense counsel, representing Amira, sought to challenge Moss's claims about the extent and cause of his injuries. In February 2002, a discovery deposition of Dr. Moser was conducted, where he provided professional insights based on his examination and treatment of Moss. In a strategic move, the defense counsel sent a letter to Dr. Moser in April 2003, just before his evidence deposition. This letter contained a detailed narrative of the medical opinions expected to be disclosed at trial, including summaries of opinions from other treating physicians and those that Dr. Moser was expected to give following his discovery deposition. This letter outlined specific views about Moss's medical condition, its causes, and the necessity of surgery, which were crucial to the defense's argument. The letter did not have consent from Moss, the patient, and was part of a legal strategy to bolster the defense's case. The defense counsel believed this approach was necessary for case preparation and did not see it as a violation of any legal or ethical standards. However, this action led to a significant legal contention, as it was argued to be an inappropriate communication, potentially influencing the testimony of a treating physician. Moss's legal team saw this as a breach of the confidentiality and fiduciary relationship between a patient and his physician, raising concerns about the integrity of the legal process and the protection of private health information.
§ Contextual Feature: Sender: Defense counsel, Sender Role: Attorney, Recipient: Dr. Richard Moser, Recipient Role: Doctor, Subject: Richard Moss, Subject Role: Patient, Type: Medical opinions and history, deposition excerpts, Purpose: To inform the physician about his expected opinions in the case, and potentially to influence the physician's future testimony, In Reply To: Plaintiff's supplemental opinion and the discovery deposition of Dr. Moser, Belief: The sender believed that this communication was necessary for case preparation and was not in violation of legal standards.
§ Compliance Conclusion: Forbid

Figure 12: A real court case sourced from CAP does not comply with HIPAA Privacy Rule.