Cloud Privacy Beyond Legal Compliance: An NLP analysis of certifiable privacy and security standards

Zachary Kilhoffer and Masooda Bashir
University of Illinois at Urbana-Champaign, School of Information Sciences

***Abstract*—By implementing standards and becoming certified, organizations can demonstrate good practices and trustworthiness. However, privacy standards are relatively immature, and the privacy research community rarely examines the individual controls of organizational standards (e.g., ISO 27017, SOC-2), which are what concretely implements privacy principles. It is also very time-consuming to monitor evolving standards, assess relevance and usefulness in a given context, and whether the effort and expense of becoming certified makes sense.**

In this paper, we propose an exploratory method leveraging a large language model (LLM) to analyze privacy documents. We created a dataset of controls (n=1,511) from all nine standards identified as certifiable, cloud relevant, and privacy relevant. We fine-tuned BERT, a popular baseline LLM, to optimize performance on privacy standards. Finally, we performed topic modeling to better understand how the standards address privacy challenges and compare to one another.

We demonstrate that controls can be grouped into 11 topics (e.g., "PII Management", "Continuous Monitoring and Auditing in Cloud"). Most standards seem to strongly emphasize the security and risk angles of privacy rather than rights and control over data. The results suggest efforts to standardize privacy practices are still nascent - more time, practice, and theoretical agreement is required before privacy standards approach the rigor of their security counterparts.

By providing our fine-tuned model, coding pipeline, and method, we demonstrate the utility of this approach to better compare and understand privacy standards and other documentation for assessment and refining.

***Index Terms*—privacy, security, certification, controls, standards**

I. Introduction

Cloud services provide powerful and scalable services that are essential to the world as we know it and many everyday services. At the same time, cloud represents a challenging environment to establish proper privacy practices.

Software engineers are often responsible for implementing specific technical and organizational privacy controls to comply with laws like GDPR (General Data Protection Regulation) and COPPA (Children's Online Privacy Protection Act). The need to comply with such laws provides extrinsic motivation to seek assurance from cloud services via certification for industry standards [17]. However, compliance with the law is a low and inadequate baseline for privacy assurances [26].

Organizations purchasing cloud services need to trust their provider; in seeking assurance of proper cloud security and privacy measures, many rely on certifiable standards. **Standards** are structured approaches outlined in formal documentation to achieve goals like compliance, interoperability, and secure operations. The key content of standards is **controls**, which are specific and verifiable organizational or technical measures. Typically standards state certain goals or principles (e.g., secure management of personal data), and a set of controls that must be fulfilled to achieve them (e.g., appropriate data retention policies, third party vetting procedures).

Standards may exceed the baseline of legal compliance and offer organizations the possibility to become certified by third party audit. Certifications for some standards are highly-regarded; ISO 27001 (from the International Organization for Standardization) is a de facto industry standard for information security systems. Standards also provide guidance for more specific contexts - like CCM (Cloud Controls Matrix) and C5 (Cloud Computing Compliance Controls Catalogue) for cloud. Certification with FedRAMP and C5 may be explicitly required to do business with the US and German governments, respectively. Previous research found that cloud customers value certification, particularly for ISO 27001 and C5 [17].

Compared to security and safety standards, there are fewer privacy standards, and these are less widely applied. The trend is towards making privacy standards more explicit, rigorous, and transparent like those in the security domain. In this process, it is essential that controls are effective in achieving privacy requirements.

The purpose of this work was to investigate controls to better understand how standards achieve privacy. We adopted a topic modeling strategy using a large language model (LLM) fine-tuned on security and privacy standards - a strategy we propose has wider applicability and promise worth exploring. We sought to answer the following main research questions :

- **RQ1**: What are the strengths and weaknesses of topic modeling in analyzing privacy and security standards?
- **RQ2**: How do privacy and security controls address privacy through technical and organizational measures?
- **RQ3**: What gaps can be identified in certifiable privacy and security standards?

We collected all privacy standards (n=9) meeting three criteria: privacy relevance; cloud relevance; and certifiability. We analyzed the texts via topic modeling – a natural language processing (NLP) method. This approach allowed us to granularly assess the standards at the level of individual controls (n=1,122), revealing shared characteristics of controls and differences between standards.

The results show that controls disproportionately emphasize the security angle of privacy, like access control, leak preven-

979-8-3503-7006-5/24/$31.00 ©2024 IEEE
DOI 10.1109/Cloud-Summit61220.2024.00020

tion, logging, and consent management. Privacy concepts like pseudonymization, control over one's data, and distinguishing personal data or PII (personally identifiable information) are absent in most standards. Controls with direct privacy relevance mostly seem to follow a risk or data control approach, rather than a rights approach.

Furthermore, standards diverge in many important ways, like the specificity of and responsibility for implementing controls. The more general controls may be applicable to more organizations, but also require more work to adapt for a given context, and complicate auditing. Our evidence supports previously identified thematic gaps [10], like mobile devices and supply chain management.

We propose that researchers should experiment with NLP techniques on standards and other privacy and security documents for developing theory, gaining new insights, and partially automating reading and comprehension tasks.

II. Related Work

Cloud computing brings many benefits, but also poses serious challenges to information privacy and security [24, 18]. Certifications are widely used to demonstrate good practices and compliance with privacy and security standards [3, 17]. In part due to overlaps between privacy and security, many standards mix the concepts; it is often unclear what individual controls aim to achieve [22].

Perhaps due to a disconnect between practitioners and the broader privacy community, few studies consider standards at the level of controls [10]. Examples include [3], which described the evolution and comprehensiveness of ISO 27001, FedRAMP, and SOC-2. Similarly, [10] analyzed FedRAMP and ISO 27001 to "improve existing standards by evaluating them with respect to known attack vectors." [8] examined SOC-2, ISO 27001, C5, and FedRAMP, concluding that three domains were under-covered: Mobile security (MOS); Interoperability and portability (IPY); and Supply chain management, transparency, and accountability (STA). [8] also identified attack vector gaps in all but SOC-2 with respect to the "treacherous twelve" identified by CSA (Cloud Security Alliance).

CSA researches and monitors IT (information technology) privacy and security standards and developed their own controls framework – CCM [7]. CCM is the largest mapping of privacy and security standards. It overviews controls from standards excluded from the present report, in addition to seven of our nine standards. However, CCM is not intended to assess controls for substantively upholding privacy and security, but moreso to simplify business decisions about seeking particular certifications [10].

Ultimately, privacy and security standards require ongoing work to remain relevant and effective. The process of improvement necessitates research to assess and compare standards, particularly as new frameworks and threats emerge [10, 8].

LLMs are known to present privacy and security issues [9], but have also shown promise for technical use cases [29] and making privacy concepts more explainable [5]. For example, [25] finetuned a pretrained LLM and achieved good results on privacy comprehension tasks. Researchers have used LLMs [16] and computational linguistics methods [1] to automate comprehension tasks for privacy policies, which are too long and numerous to realistically read [14]. One valuable use case is generating "privacy nutrition labels" from privacy policies [16]. Researchers like [6] have also created benchmarks to evaluate LLMs on privacy language understanding.

Our contribution is twofold. First, we make a methodological contribution. We provide a pipeline, code, and a LLM model that we fine-tuned with a domain-specific corpus and a small, human-labeled dataset of privacy and security controls. This represents a rather novel use case for LLMs in the privacy domain. Given the rapid development of LLMs and their potential to wholly or partially automate tasks, it is important that researchers continue to propose and test new use cases. Second, we comprehensively compare the controls of certifiable cloud privacy standards. The results are primarily intended to inform practitioners tasked with updating and improving cloud privacy frameworks. Additionally, CSPs (cloud service providers) and customers could benefit from a better understanding of privacy standards, helping them to evaluate options and judge the value of a given certification.

III. Methodology

We primarily used a topic modeling pipeline. Topic modeling is a form of unsupervised machine learning (ML) that allows researchers to investigate themes or topics within or between documents without any prior knowledge. While especially valuable in exploratory work, topic modeling does not have a ground truth or labels by which to benchmark results or assess validity and reliability. Topic modeling strongly benefits from a human-in-the-loop with domain expertise [2], and the quality of results is inherently somewhat subjective [15].

New NLP techniques have made topic modeling easier and more robust, and many studies have shown that minimal fine-tuning can vastly improve results [28, 20]. Fine-tuning the language model helps it better understand and generate text specific to a particular field or topic, which improves the model's accuracy and relevance when dealing with content from that domain. We used labeled data as well as unstructured, domain-specific text for fine-tuning.

A. Data Collection

We first established the inclusion criteria for documents of interest. Our goal was to select documents for which the analysis would be most useful to the privacy community, and documents that are reasonably similar, and thus fair to compare. We thus analyzed all privacy standards (also called privacy frameworks, guidelines, etc.) that met three criteria: (1) privacy relevance, (2) cloud relevance, and (3) certifiability.

Privacy relevance and cloud relevance are somewhat subjective, and we identified privacy and cloud relevant standards in cooperation with industry experts. Certifiability means organizations can attain certification via third party audit for compliance with a standard. We would expect certifiable standards to

TABLE I: Overview of Assessed Standards

Standard	Creator	Privacy Concepts	Original Release	Version	Controls From or Combined With	Intended Audience
C5	German gov.	Personal data, sensitive data, data protection	2016	C5:2020	ISO 27001, 27002, 27017; BSI - IT-Grundschutz-Kompendium v2; CCM	CSPs, CSP auditors, cloud customers
CCM	US non-profit	Personal data, sensitive data	2010	CCM 4.0.10	Unclear	CSPs, cloud customers, Auditors
EU CoC	European Commission	Data protection, personal data	2017	2.11	GDPR	CSPs, 3PAOs engaged by CSP or customers
FedRAMP	US gov.	PII	2011	Rev. 5 (high)	NIST SP 800-53	CSPs who work with gov., 3PAOs
SOC-2	US non-profit	Personal data, sensitive personal data	2010	2017 TSC RPF 2022	ISO 27001, COSO Internal Control Integrated Framework	Organizations providing services, customers
ISO 27002	Swiss NGO	PII; personal, confidential, sensitive data	2013	ISO 27001:2022E	N/A	Organizations of all types and sizes
ISO 27017	Swiss NGO	PII, sensitive data	2015	ISO 27017:2015	ISO 27002	CSP, cloud customers
ISO 27018	Swiss NGO	PII, personal data, sensitive personal data	2014	ISO 27017:2019E	ISO 27002	Public CSP acting as a PII processor
ISO 27701	Swiss NGO	PII, PIMS	2019	ISO 27001:2019E	ISO 27001, 27002	PII controllers, PII processors

contain rigorous and auditable privacy controls. While certifiability excluded influential standards like the NIST (National Institute of Standards and Technology) Privacy Framework, we needed comparable documents for fair comparison.

Having determined the inclusion criteria, we searched for privacy documents meeting our criteria and acquired the most recent versions. The documents are C5, CCM, EU COC, FedRAMP (High baseline), ISO 27001, ISO 27017, ISO 27018, ISO 27701, and SOC-2 (see Table I for overview).

The standards selected, and their controls, differ in ways that may complicate topic modeling. Examples include what type of system (cloud, other); the parties responsible based on cloud service (e.g., SaaS – software as a service – implies CSP responsible for security, IaaS – infrastructure as a service – implies tenant and provider share responsibility) [13, 19]; whether the control is intended for senior management, business or process managers, or operations/implementation level; legal requirements imposed based on location or sector (e.g., California Online Privacy Protection Act, COPPA for children), etc. These differences mean we must be careful to consider what unit of meaning is encapsulated in each topic emerging from the topic modeling exercise.

B. Data cleaning

Only PDF copies of some standards (notably ISO's) are published, with unselectable text, whereas others like FedRAMP have machine readable formats available. We needed to use optical character recognition (OCR) techniques to extract text from the PDFs since simple PDF libraries could not interpret it. However, OCR remains error prone [12].

There are many strategies to minimize but not eliminate OCR errors, necessitating a correction step in the pipeline. We manually examined results and used regular expressions to correct frequent errors. Still, a great deal of manual work would have been required to fix all errors and bring all data to a high standard of quality. We therefore relied on a LLM to spot and correct OCR errors. This meant querying the OpenAI API with text converted from PDFs and a standardized prompt. This approach has been used in several recent papers and shown good results [23]. The prompt essentially reads:

> I will provide you text from cybersecurity and privacy documentation like ISO 27017, FedRAMP, etc. The text is the result of imperfect optical character recognition (OCR). Your job is to return the same text with OCR errors corrected.

What is a control? Next, we needed to filter and transform the data such that each row represented a single control. This was trivial for documents like FedRAMP that started as CSVs, but for those starting as PDFs, a significant amount of coding and manual work was required.

Despite different naming conventions, it is mostly obvious what constitutes a control, and where one begins and ends. For example, SOC-2 has a table of "trust services criteria" (essentially objectives) and "related points of focus" (essentially controls). The most recent version of C5 technically has no controls, as the name changed from Cloud Computing Compliance *Controls* Catalog to Cloud Computing Compliance *Criteria* Catalog; regardless, they are all controls for our purposes.

Some human judgment was still required to say precisely

where one control ends and another begins, especially when control coding may suggest a control with multiple components. We identified controls as the smallest text units identified with a unique name or code, and we opted to be inclusive with respect to what text is part of a control. For example, some controls contain "additional guidance" or similar, which we simply included in the control text. Thus, a single control in our data often had multiple paragraphs or bullet points.

C. Describing the data

These steps resulted in a dataset of 1,561 rows where each row represents a unique control. Columns contain metadata such as the standard name, page number, control code, control title, control text, etc.

We did not use all 1,561 controls for topic modeling because many were not *substantive*, but merely *referential*. Referential controls typically state something like "The guidance from Document X Section Y applies." While such controls refer to substantive controls in other documents, the texts themselves would simply add undesirable noise to the data (e.g., additional standards' names and acronyms). Thus, we removed 439 controls identified as "strictly referential" as follows: filter rows matching rules (under 250 characters in length, matches regex like "the requirement stated in .* applies"), then manually checking results. The final dataset we used for topic modeling has 1,122 substantive controls.

D. Topic modeling pipeline.

We used a Python library called BERTopic [11]. BERTopic uses a six-step algorithm for topic modeling (Fig. 1 shows steps with defaults methods) and often yields solid results with default settings. However, each step is modular (e.g., HDBScan and K-nearest neighbor are interchangeable for clustering) to allow flexibility for a given corpus and goal. BERTopic also allows us to easily fine-tune representations of clusters (note: this is separate from fine-tuning the LLM), which is very helpful for humans to understand what a given topic represents.

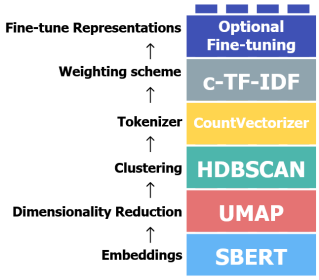


Fig. 1: BERTopic Algorithm

Several methodological choices are noteworthy. The LLM creates embeddings (vectors of numbers) from control texts, and the embeddings themselves are what we perform cluster analysis on. We opted for a very popular general use case LLM: BERT (specifically BERT Base Cased), which embeds texts up to 512 tokens long. The choice of a cased model (lowercase and uppercase are different) is mostly because of the prevalence of acronyms in the data. While BERT no longer represents the state-of-the-art, it remains a very useful baseline and achieves good performance on many tasks. Larger (more parameters) and more sophisticated LLMs can generate "better" embeddings containing more data, but dimensionality reduction will then reduce the size of embeddings anyway, so the value added is questionable. We found the larger BERT models actually harmed performance with noisier results.

Fine-tuning. For fine-tuning, we first fed the LLM a domain-specific corpus – essentially all substantive controls. Since our goal was to create meaningful labels for controls, we further fine-tuned with a sample of controls we manually labeled.

This required potential labels that we might expect controls to match. We experimented with options like the Fair Information Practice Principles (FIPPs), Comprehensive Privacy Criteria Framework (C2P2) [27], etc. Considering our particular data and preliminary results, we decided on a framework that included a moderate number of privacy and security concepts. This was necessary because so many of the controls turned out to be security-oriented, and thus poor matches for common privacy concepts like consent management and confidentiality. The framework we used is called Cybersecurity & Data Privacy by Design Principles [21] which contains 33 labels. With this framework, the first author and two other researchers each labeled a random selection of 30 substantive controls. The Fleiss' Kappa score was initially .26, indicating fair agreement. The authors then discussed results and were allowed to change their answers. The reconciled results showed a score of 0.71, indicating substantial agreement. We used the data for which all or 2/3 researchers agreed on the label (n=25) for fine-tuning.

Clustering. All clustering methods have pros and cons to consider. We chose HDBScan primarily because it creates a hierarchical classification of topics. This method allows researchers flexibility to merge lower order (more granular) topics based on the desired specificity, and whether a given distinction is conceptually useful. Also, HDBScan assigns each data point (controls, in our case) a set of probabilities for multiple topics. This can be very useful when expecting conceptual overlap between topics, as we would with complex technical texts about privacy and security.

Topic representation is crucial for interpretability. To understand what a topic represents and assign it a descriptive name, we first look at the tokens, words, or n-grams with the strongest associations (i.e., proximity in a high-dimensional space) to the topic's centroid, which is basically the average of text embeddings in a topic cluster. However, associated tokens do not always make sense to humans.

Fine-tuning the topic representation helps us create human-interpretable topic names. We performed two steps to this end. First, we generated higher-quality token representations of each topic (fewer stop words, greater coherence) using KeyBERTInspired – a model specifically for fine-tuning topic representations. Second, we used GPT-4 via OpenAI's API. For each topic, we passed representative examples of control texts, the five most associated n-grams, and the following instruction:

> I have a topic that contains the following documents: [DOCUMENTS]. The topic is described by the following keywords: [KEYWORDS]. Based on the information above, extract a short but highly

descriptive topic label of at most 5 words. Make sure it is in the following format:
topic: <topic label>

We then determined final topic names with the associated n-grams, GPT-4's suggestion, and representative controls.

Merging topics for parsimony. At this point, there is no need to continue provided the topics are useful. However, topic modeling often results in multiple highly similar topics [4], and we may want a more manageable number of topics for better understandability. To this end, there are a few options.

We must set the minimum cluster size for HDBScan, so we could simply raise this value to reach the desired number of topics. However, it is hard to assess within-cluster similarity if the topics are too few and too general. We therefore began with a lower minimum cluster size – thus higher number of topics – to ensure that results were good at a granular level. After experimentation, the best results came from initially setting a minimum cluster size=5, which initially yielded 49 topics.

Next, we manually inspected and merged topics to achieve more parsimonious and interpretable results (see Figure 2). This mostly involved looking at two adjacent topics and considering if their distinction is useful. We also looked at data visualizations of the topic hierarchy, as more distance between nodes (x-axis) indicates more conceptual difference (see Table II). After merging topics, the centroid is recalculated and new associated n-grams produced. To explain a reasonable number of topics, we manually inspected and merged, finishing with 11 topics.

IV. Results

After merging, we achieved a parsimonious and interpretable overview with 11 topics, summarized below and in Table II.

A. Topics

1. Audit, Governance and Supply Chain Management is mostly about how audits should be conducted to assure proper data governance and supply chain management. Controls in this topic mostly represent CCM, which is sensible as it is an auditing standard.

2. CSP Responsibilities and Guidance refers to controls that are very specific to cloud – typically the way CSPs should handle personal data, such as offering customers way to control their own data. Most controls here are from the EU COC and ISO 27017, which are cloud-specific standards and more oriented towards privacy than security.

3. Continuous Monitoring and Auditing in Cloud represents controls that secure against common threats in cloud environments. For this topic, most controls are from C5, which heavily focuses on cloud security. In fact, all of C5's controls fell into this category. The difference between topics 1 and 3 is that 3 is more cloud specific and emphasizes the continuity of auditing. All standards but ISO 27017 have a control here.

4. Control Management and Communication mostly represents controls focused on how organizations should handle security and communicate. These controls seem mostly applicable to SaaS organizations, which is sensible given that the topic represents 82% of SOC-2 controls.

5. Cryptography & Key Management is straightforward. Controls in this grouping are about secure key practices, encryption, and similar, but not access control (managing who can access and change resources within a network or computer system) per se.

6. Identification, Authentication, & Verification is a more security-focused topic. This involves methods such as passwords, biometrics, and multi-factor authentication; it emphasizes confirming who is accessing the system or data. These controls are mostly from FedRAMP.

7. PII Management is very privacy-focused and rather self-explanatory. Of note, three ISO standards are included in this topic, while six standards do not seem to directly address the management of PII management (beyond security controls).

8. Policies and Procedure Management involves the development, documentation, and dissemination of security policies and procedures to comply with FedRAMP. This topic only contains FedRAMP controls, which cover risk assessment, incident response planning, assigning responsibilities to appropriate individuals, and providing training for employees.

9. Security Guidance for PII in Cloud involves specific measures for securing PII, particularly in cloud settings. Example controls include those related to temporary and emergency account management, disposal of temporary files created from PII, and specifying minimum technical and organizational measures for data processing.

10. System Boundaries and Access Control involves isolating environments from one another, and access control. Many controls within this topic detail the mechanisms to authenticate users or authorize access based on user permissions.

11. Identity and Privileged Access Management relates to access control, but goes further to control the management of user identities and privileged access. For example, this includes provisioning (creating and managing user accounts) and deprovisioning (disabling or deleting accounts). Additionally, controls concern security practices for administrators or system operators with elevated permissions and access rights, and how these should be monitored and restricted.

B. Interpreting results

These 11 topics should be interpreted as high-level descriptions of similar controls. Assessing topic modeling quality involves: (1) similarity within clusters; (2) difference between clusters; and (3) accuracy and usefulness of interpretation. Still, assessment is challenging – without "ground truth" labels for benchmarking, quality is somewhat subjective. Additionally, similarity within cluster trades off with the parsimony of the topic model; by merging topics, we grouped different sets of controls for the sake of explainability. We sought to balance granularity and accuracy with explainability and interpretability.

V. Discussion

This paper's contributions are (1) demonstrating a novel application of LLMs in the field of privacy and security, and (2) presenting topic modeling results for all nine certifiable cloud privacy standards. We contend that our study is proof

TABLE II: Topic Modeling Results: Count of controls per topic (normalized by controls per standard)

	Audit, Governance & Supply Chain Management	CSP Responsibilities & Guidance	Continuous Monitoring & Auditing in Cloud	Control Management & Communication	Crytography & Key Management	Identification, Authentication, & Verification	PII Management	Policies & Procedures Management	Security Guidance for PII in Cloud	System Boundaries & Access Control	Identity & Privileged Access Management	**Total**
C5			121 (100.0%)									121
CCM	142 (72.1%)	3 (1.5%)	6 (3.0%)	2 (1.0%)	19 (9.6%)	2 (1.0%)			1 (0.5%)	12 (6.1%)	10 (5.1%)	197
EU COC	1 (1.6%)	57 (90.5%)	3 (4.8%)		1 (1.6%)				1 (1.6%)			63
FedRAMP	17 (4.1%)		15 (3.7%)		15 (3.7%)	29 (7.1%)		33 (8.0%)	3 (0.7%)	119 (29.0%)	179 (43.7%)	410
ISO 27002	2 (4.9%)		1 (2.4%)				8 (19.5%)		29 (70.7%)	1 (2.4%)		41
ISO 27017		44 (93.6%)							1 (2.1%)		2 (4.3%)	47
ISO 27018	2 (4.9%)		1 (2.4%)				9 (22.0%)		28 (68.3%)	1 (2.4%)		41
ISO 27701	1 (0.7%)		1 (0.7%)				92 (65.2%)		47 (33.3%)			141
SOC-2			3 (4.9%)	50 (82.0%)						6 (9.8%)	2 (3.3%)	61
Sums	165	104	151	52	35	31	109	33	110	139	193	1122

of concept that LLMs can partially automate the work of comparing and assessing privacy standards, though we also used human evaluation at each step to ensure quality.

RQ1: What are the strengths and weaknesses of topic modeling in analyzing privacy and security standards? The first challenge we encountered was a lack of objective ways to assess how good results are. We carefully inspected results at each stage to check that groupings were logical, capturing similar controls. However, clusters might not capture the desired attributes. We were primarily interested in understanding *how controls accomplish privacy*, but some initial attempts seemed to cluster controls naively, like batching all controls that contain "CSP" regardless of context. Furthermore, hyperparameter adjustments like the minimum number of items per cluster have huge effects on cluster size and quality.

We have two suggestions. First, follow best practices for topic modeling, as we have tried, and rely on domain expertise and intuition to gauge if results are logical and useful, or noisy and arbitrary. Second, consider hierarchical clustering. One can start with many small clusters ordered hierarchically, then merge topics until reaching the desired level of generalization. There is no strictly correct level of specificity, but different use cases may benefit from more or less granularity.

Generally, we found that topic modeling makes it fairly easy to assess and compare controls across standards. Especially when more and longer documents are of interest, this could be quite valuable. For example, one could gather all controls relevant to PII without much manual work.

Finally, we found that BERT did not produce good (i.e., appearing logical) results until we had finetuned it. This reflects growing literature about the modest work but large gains from LLM finetuning. Also, BERT is becoming dated, and more recent LLMs are a better starting point for this pipeline.

RQ2: How do privacy and security controls address privacy through technical and organizational measures? We identified how security and privacy controls work, as denoted by initial 49 topics (omitted for length), and the more general topics presented above. As [22] observed, standards seem to mix privacy and security concepts. This, combined with the lack of privacy specific language in documents like C5 and SOC-2, indicates that standards tend towards a security approach.

RQ3: What gaps can be identified in certifiable privacy and security standards? Our findings support [10] which found that controls supporting mobile security are lacking in current cloud standards, though we found several like FedRAMP high have great detail in supply chain management, transparency, and accountability. This could indicate that previous research did not consider all relevant standards, or standards have improved over time. While we found mobile security measures are lacking, we hesitate to propose further gaps based on our analysis due to methodological limitations (see below).

Implications. First, we suggest that LLMs have many untapped applications in privacy, security, standards, and compliance. Second, we find that most controls heavily lean towards a security-centric notion of privacy. For example, access control is ubiquitous across documents, but distinctions between private/personal data and other data are not. The results may suggest that developing a set of broadly-accepted privacy controls remains very difficult. We propose that privacy standards should strive for the rigor, gravitas, and acceptance of security standards. However, further work is required to improve the individual controls, and broader agreement is required on what doing privacy means beyond compliance.

Limitations and future research. This work has limitations. Not all the standards considered are strictly for privacy; some mostly cover security. However, the messy boundary of privacy and security is a limitation imposed by the world as it currently is, and we erred on the side of including standards that might be relevant. While we have tried to select standards conducive to fair comparison, they vary significantly, like focusing on a technical or managerial level.

We only analyzed substantive controls, which understates ISO standard coverage, and did not analyze results for accuracy with available crosswalks (like CCM's). Therefore, we feel our results are unsuited to make strong claims about gaps.

We propose further work on fine-tuning LLMs and testing them for specific privacy applications. For example, if an LLM were fed (a) a company's privacy policies and obligations; (b) the text of a standard; and (c) carefully engineered prompts, it may produce a useful *first attempt* at adapting the controls to the company's specific context. We will build on these results in several ways: using a newer LLM than BERT, validating the prompts used to instruct an LLM, using the actual texts referential controls point to (rather than omitting them), and leveraging control crosswalks. While crosswalks may contain mistakes or errata [10], we consider them useful for applications

like training LLMs to spot similar controls and evaluate topic modeling results.

Overall, we advocate something like our approach for difficult and time-consuming tasks involving privacy and security documentation. LLMs likely have many more useful domain applications waiting to be discovered.

VI. Acknowledgements

This research benefited from generous funding from Cisco.

References

[1] Wasi Uddin Ahmad et al. *Intent Classification and Slot Filling for Privacy Policies*. June 4, 2021.

[2] David Andrzejewski, Xiaojin Zhu, and Mark Craven. "Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors". In: June 14, 2009, pp. 25–32.

[3] Masooda Bashir, Carlo Di Giulio, and Charles A. Kamhoua. "Certifications Past and Future: A Future Model for Assigning Certifications That Incorporate Lessons Learned from Past Practices". In: July 18, 2018.

[4] J. Boyd-Graber, D Mimno, and D Newman. "Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements". In: 2015.

[5] Yufan Chen, Arjun Arunasalam, and Z. Berkay Celik. "Can Large Language Models Provide Security & Privacy Advice? Measuring the Ability of LLMs to Refute Misconceptions". In: ACSAC '23. Dec. 4, 2023, pp. 366–378.

[6] Jianfeng Chi et al. *PLUE: Language Understanding Evaluation Benchmark for Privacy Policies in English*. May 12, 2023.

[7] Cloud Security Alliance. *Cloud Controls Matrix (CCM)*. CSA. 2024.

[8] "Cloud Security Certifications: A Comparison to Improve Cloud Service Provider Security". In: Mar. 22, 2017, pp. 1–12.

[9] Erik Derner and Kristina Batistič. *Beyond the Safeguards: Exploring the Security Risks of ChatGPT*. May 13, 2023.

[10] Carlo Di Giulio et al. "Cloud Standards in Comparison: Are New Security Frameworks Improving Cloud Security?" In: June 2017, pp. 50–57.

[11] Maarten Grootendorst. *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure*. Mar. 11, 2022.

[12] Ming Jiang et al. "Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book Excerpts". In: Nov. 2021.

[13] Tim Mather, Subra Kumaraswamy, and Shahed Latif. *Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance*. Sept. 4, 2009. 338 pp. Google Books: BHazecOuDLYC.

[14] Aleecia M. McDonald and Lorrie Faith Cranor. "The Cost of Reading Privacy Policies". In: *I/S: A Journal of Law and Policy for the Information Society* 4.3 (2008–2009), pp. 543–568.

[15] John W. Mohr and Petko Bogdanov. "Introduction—Topic Models: What They Are and Why They Matter". In: *Poetics*. Topic Models and the Cultural Sciences 41.6 (Dec. 1, 2013), pp. 545–569.

[16] Shidong Pan et al. *Toward the Cure of Privacy Policy Reading Phobia: Automated Generation of Privacy Nutrition Labels From Privacy Policies*. June 19, 2023.

[17] Sebastian Pape and Jelena Stankovic. "An Insight into Decisive Factors in Cloud Provider Selection with a Focus on Security". In: Cham, 2020, pp. 287–306.

[18] Charith Perera et al. "Big Data Privacy in the Internet of Things Era". In: *IT professional* 17.3 (2015), pp. 32–39.

[19] C. Saravanakumar and C. Arun. "Survey on Interoperability, Security, Trust, Privacy Standardization of Cloud Computing". In: *2014 International Conference on Contemporary Computing and Informatics (IC3I)*. 2014 International Conference on Contemporary Computing and Informatics (IC3I). Nov. 2014, pp. 977–982. DOI: 10.1109/IC3I.2014.7019735.

[20] Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. *Fine-Tuned Language Models Are Continual Learners*. Oct. 29, 2022.

[21] Secure Controls Framework. *Cybersecurity & Data Privacy by Design Principles*. Mar. 2023.

[22] Tanusree Sharma et al. "Towards Inclusive Privacy Protections in the Cloud". In: 2020, pp. 337–359.

[23] Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. "BART for Post-Correction of OCR Newspaper Text". In: 2021, pp. 284–290.

[24] Ali Sunyaev and Stephan Schneider. "Cloud Services Certification". In: 56.2 (2013), pp. 33–36.

[25] Chenhao Tang et al. *PolicyGPT: Automated Analysis of Privacy Policies with Large Language Models*. Sept. 18, 2023.

[26] Ari Ezra Waldman. *Privacy Law's False Promise*. SSRN Scholarly Paper. Rochester, NY, Dec. 2019.

[27] Tian Wang, Carol Mullins Hayes, and Masooda Bashir. "Developing a Framework of Comprehensive Criteria for Privacy Protections". In: Lecture Notes in Networks and Systems (2022), pp. 905–918.

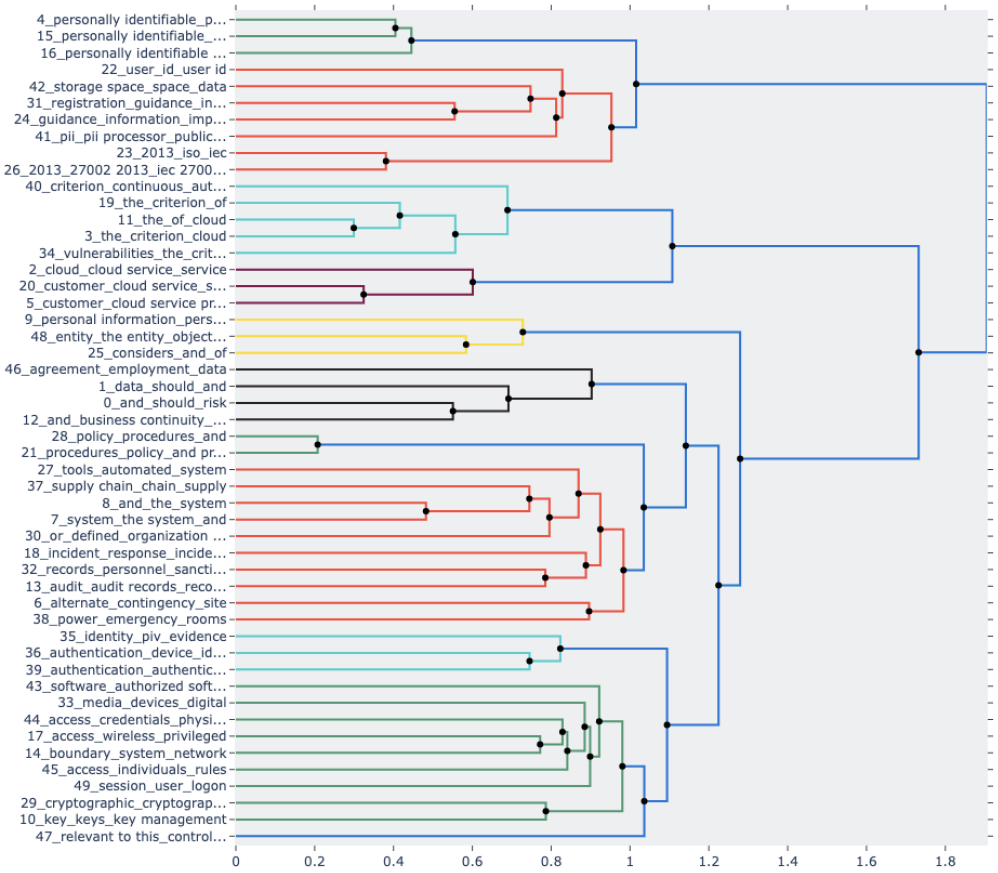
[28] Jason Wei et al. *Finetuned Language Models Are Zero-Shot Learners*. Feb. 8, 2022.

[29] Yifan Yao et al. "A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly". In: *High-Confidence Computing* 4.2 (June 1, 2024), p. 100211.

VII. Appendix

Code and fine-tuned model available at:
https://github.com/zkilhoffer/privacy-standard-topic-modeling

Fig. 2: Hierarchical Clustering Results: Initial 49 topics



Note 1: The 49 labels on the left show the initial topic representations before their optional fine-tuning. Many stop words make human interpretation difficult – for example, the most associated n-gram for topic 25 is "considers and of". Fine-tuning the topic representations results in more informative and human-comprehensible labels.

Note 2: Hierachical clustering allows us to merge some of the original 49 topics for a more parsimonious overview. If we choose a spot on the x-axis and draw a vertical line up, then count the number of horizontal lines crossed, that would be *the number of topics at a given level of granularity*. Starting at 0 on the x-axis and moving up, we would cross 49 lines and have 49 topics. This is the original, most granular result. Starting from 1 on the x-axis, we would cross about ten lines, and thus have ten topics instead of the original 49 – a fairly high-level topic representation. Manually merging nodes is also possible, after which the centroid is recalculated with the new, combined set of controls.