PrivaT5: A Generative Language Model for Privacy Policies

Mohammad Al Zoubi, Santosh T.Y.S.S,
Edgar Ricardo Chavez Rosas, Matthias Grabmair
School of Computation, Information, and Technology
Technical University of Munich, Germany
{mohammad.al-zoubi, santosh.tokala, matthias.grabmair}@tum.de
e.ricardo.chavez@hotmail.com

Abstract

In the era of of digital privacy, users often neglect to read privacy policies due to their complexity. To bridge this gap, NLP models have emerged to assist in understanding privacy policies. While recent generative language models like BART and T5 have shown prowess in text generation and discriminative tasks being framed as generative ones, their application to privacy policy domain tasks remains unexplored. To address that, we introduce PrivaT5, a T5-based model that is further pre-trained on privacy policy text. We evaluate PrivaT5 over a diverse privacy policy related tasks and notice its superior performance over T5, showing the utility of continued domain-specific pretraining. Our results also highlight challenges faced by these generative models in complex structured output label space, especially in sequence tagging tasks, where they fall short compared to lighter encoder-only models.[1]

1 Introduction

Privacy policies outline how companies collect, use, share and manage user data on their services or applications. They are governed by a framework of notice and choice in many jurisdictions (Landesberg et al., 1998), requiring website operators to post a notice about how they gather and process users' information. Users then decide whether to accept or abstain from using the website or service. However, the effectiveness of this framework, even enshrined in regulations like GDPR, relies on users comprehending these policies, which is often not the case due to their length, legal complexity and reasoning over vagueness and ambiguity (Gluck et al., 2016; Reidenberg et al., 2016; FTC).

Moreover, the prevalence of data surveillance and misuse, exemplified by scandals involving companies like Facebook and Cambridge Analytica (Cadwalladr and Graham-Harrison, 2018), underscores the critical nature of privacy concerns in the digital era. This scenario provides an ideal context for advancements in NLP to provide users with tools to understand policy content and address their privacy inquiries effectively. Harnessing NLP advancements would benefit not only individuals but also assist companies in ensuring compliance and regulators in enforcing it across diverse software products and services (Ravichander et al., 2021). It's important to note that privacy policies stand apart from closely related domains, like legal texts (Shankar et al., 2023) which are tailored for domain experts. Instead, privacy policies, as legal documents with legal implications, are generally composed by experts, yet intended to be comprehensible by everyday users.

There have been significant research effort devoted to automate the analysis of privacy policies under Usable Privacy Project (Sadeh et al., 2013). Some works include identification of policy segments commenting on specific data practices (Wilson et al., 2016), compliance analysis (Zimmeck et al., 2019), extraction of opt-out choices (Sathyendra et al., 2017; Bannihatti Kumar et al., 2020), text alignment (Ramanath et al., 2014), vague sentence detection (Lebanoff and Liu, 2018), question answering (QA) (Ahmad et al., 2020; Ravichander, 2019; Harkous et al., 2018), summarization (Keymanesh et al., 2020; Zaeem et al., 2018), readability analysis (Meiselwitz, 2013; Massey et al., 2013) and fine-grained structured information (Hosseini et al., 2020; Le et al., 2021; Bui et al., 2021).

Earlier works focusing on privacy policies utilized extensive feature engineering (Wilson et al., 2016; Sathyendra et al., 2017; Zimmeck et al., 2019), domain-specific word embeddings (Kumar et al., 2019) and with the rise of pre-trained models like BERT, the pretrain-then-finetune approach has gained prominence (Mousavi Nejad et al., 2020; Ravichander, 2019; Ahmad et al., 2020). More-

[1] Our pre-trained PrivaT5 models are available at https://github.com/TUMLegalTech/PrivaT5.

Proceedings of the Fifth Workshop on Privacy in Natural Language Processing, pages 159–169

over, Gururangan et al. 2020 emphasized that further continuing the pre-training of language models on domain-specific corpora can further elevate model performance in tasks specific to that domain. This, coupled with the availability of extensive privacy policy corpora (Srinath et al., 2021; Amos et al., 2021), has paved the way for developing privBERT (Srinath et al., 2021). This model excels in privacy language understanding tasks, as evidenced by its performance on constructed benchmarks designed in the privacy domain, such as PrivacyGLUE (Shankar et al., 2023) and PLUE (Chi et al., 2023).

More recently, there has been growing interest in generative language models, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), due to their inherent effectiveness in natural language generation tasks like summarization, question answering, and simplification. These generative models enable a unified approach to both discriminative and generative tasks by framing various non-generative tasks in a text-to-text format. However, the privacy domain lacks dedicated generative models and the exploration of casting non-generative tasks into a generative format remains uncharted. To address this gap, we embark on pre-training T5 models on the Privaseer corpus, resulting in various PrivaT5 variants across small (60M parameters), base (220M parameters) and large (770M parameters) sizes. We systematically evaluate the performance of both PrivaT5 and T5 on a range of privacy policy-related tasks to assess their capabilities along the axes of model size and pre-training corpus. Our results demonstrate the impact of pre-training using domain related corpora on the downstream task performance while highlighting the challenges of generative models dealing with structured output in information extraction tasks.

2 PrivacyT5

T5 is an encoder-decoder model initially pre-trained in an unsupervised manner on the C4 corpus (Raffel et al., 2020). This pre-training involves replacing 15% of the tokens with sentinel tokens in a denoising objective, with consecutive tokens marked for removal being replaced by a single sentinel token. The resulting corrupted text serves as input to the model to predict the masked-out tokens. Then the model is further fine-tuned using supervised training on various downstream tasks, including those from the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks, casting them into text-to-text format for training.

To pre-train the PrivaT5 models, we initialize the model with T5 [2] and continue pre-training with the PrivaSeer Corpus (Srinath et al., 2021), which encompasses 1,005,380 privacy policies originating from 995,475 distinct web domains with prominent ones like .com, .org, and .net comprising significant proportions of the corpus at 63%, 5%, and 3%, respectively. We pre-train small (60M), base (220M) and large (770M) versions of T5 to obtain privaT5 models of three sizes. Detailed hyperparameters related to pre-training can be found in App. D.

3 Experiments

We evaluate the models on the following privacy policy related downstream tasks. App. A and B describe dataset splits with their label space and illustrative instances respectively.

OPP115 (Wilson et al., 2016; Mousavi Nejad et al., 2020) consists of 3432 sentences from 115 online privacy policies annotated with one or more privacy practices from ten categories to aid compliance analysis, leading to a multi-label classification.

PI-Extract (Bui et al., 2021) focuses on extracting token spans representing data-related entities such as collected, not collected, not shared, and shared, akin to Named Entity Recognition. This dataset comprises 4064 sentences extracted from 30 privacy policy documents. Notably, the entities of various types may overlap, leading to a token-level multi-label classification approach.

PolicyDetection (Amos et al., 2021) includes 1301 documents focusing on binary classification, categorizing as either privacy policies related or not.

PolicyIE (Le et al., 2021) consists of 5250 sentences, each labelled with a privacy practice intent label (referred to as task IE-A), and the word spans annotated with a slot label (referred to as task IE-B) derived from 31 privacy policies of websites and mobile applications. IE-A has 5 intent classes and IE-B has 18 slot labels, categorized into 14 type-I slots for privacy practice participants and 4 type-II slots for details like purposes and conditions. Note that type-I and type-II slot values in IE-B can overlap resulting into a joint multi-label classification, while IE-A is a multi-class classification task.

PrivacyQA (Ravichander, 2019) is comprised of 1750 questions related to the privacy policies of mobile applications. This task is framed as binary

[2] https://huggingface.co/docs/transformers/model_doc/t5

	OPP 115	PI Extract	Policy Detection	Policy IE-A	Policy IE-B	Privacy QA	Policy QA	Policy Summ
STL								
T5 (Small)	77.03	52.34	83.35	68.74	44.24	47.24	18.15	0.445/0.253/0.433
PrivaT5 (Small)	77.35	60.48	84.16	70.88	46.23	51.13	20.46	0.462/0.262/0.450
T5 (Base)	79.12	62.54	87.52	73.45	46.17	48.46	22.14	0.539/0.350/0.526
PrivaT5 (Base)	80.53	61.98	86.65	77.74	48.29	56.13	24.16	0.563/0.372/0.549
T5 (Large)	81.58	63.97	88.78	76.28	48.28	56.28	25.17	0.557/0.362/0.544
PrivaT5 (Large)	81.49	66.34	88.71	78.09	51.76	63.38	27.14	0.575/0.388/0.565
BERT	77.82	60.25	85.21	71.87	50.18	53.24	28.23	-
LegalBERT	78.34	58.98	86.13	72.28	51.27	53.36	27.37	-
PrivBERT	81.56	63.36	87.24	75.14	54.28	55.32	31.14	-
MTL								
T5 (Small)	75.34	54.29	81.14	72.86	45.12	45.20	17.19	0.331/0.178/0.318
PrivaT5 (Small)	76.28	60.87	84.22	73.34	46.78	47.72	18.16	0.349/0.192/0.336
T5 (Base)	77.02	56.78	86.29	76.12	46.22	48.12	19.46	0.463/0.285/0.451
PrivaT5 (Base)	77.24	62.83	86.12	76.68	47.28	50.14	20.48	0.484/0.321/0.471
T5 (Large)	77.84	60.04	86.88	77.28	46.78	49.87	22.66	0.473/0.278/0.461
PrivaT5 (Large)	78.82	64.24	87.43	78.88	47.62	51.14	24.22	0.508/0.334/0.492

Table 1: Performance comparison over different downstream tasks. ROUGE-1/2/L scores, Exact Match are reported for PolicySumm and PolicyQA respectively and Macro-F1 scores are reported for rest of the tasks.

relevance prediction, where the objective is to determine whether a given sentence from a privacy policy is relevant to a specific question.

PolicyQA (Ahmad et al., 2020) contains 25,017 reading comprehension style questions curated from 115 website privacy policies. Unlike PrivacyQA, which focuses on sentence-level answers from policy documents, PolicyQA adopts a setup similar to SQUAD (Rajpurkar et al., 2016), where it requires a shorter text span as the answer given the corresponding policy document and question.

PolicySumm (Kumar et al., 2022; Gopinath et al., 2020) consists of 24000 section body, title pairs from privacy policies where the task involves generating section title given the content of section.

Evaluation Metrics We report macro-F1 for all the classification tasks such as OPP115, PolicyDetection, PolicyIE-A, PrivacyQA. For PI-Extract and PolicyIE-B, we compute the macro-F1 scores for each entity obtained from token-level labels. For PolicyQA, we report the exact match which measures percentage of predictions that match any one of the ground truth answers exactly. For PolicySumm, we report ROUGE-1,2 and L scores.

Implementation Details We convert each of the task into text-to-text format where the model produces output in the form of text. The model is directly trained with a maximum likelihood objective using teacher forcing, regardless of the task, unifying the pre-training and fine-tuning objective. In case of multi-class/binary classification problem (such as PolicyDetection, PolicyIE-A, PrivacyQA), the output label is verbalized into text format (such as 'Policy' and 'Not a Policy' in case of PolicyDetection). In case of multi-label classification (such as OPP115), we verbalize the class labels into texts and concatenate the multiple labels using a delimiter. For sequence tagging (NER kind of task such as PolicyIE-B and PI-Extract), we use 'Sentinel + Tag' strategy described in Raman et al. 2022, where the sentinel tokens $< extra\_id\_0 >$, $< extra\_id\_1 >$ etc are incorporated before each token wile feeding input to the model and the output is produced by generating respective sentinel token along with its output tag. For PrivacyQA and PolicySumm, we allow the model to generate the free-form text. Text-to-text transformations on illustrative examples are provided in Appendix C. We assess models performance on each of the task independently, referred to as *Single Task Learning (STL)*, by initializing with {T5/PrivaT5}-{Small/Base/Large} version and fine-tuning it on the task-specific training data. Further, we also assess the *Multi Task Learning (MTL)* ability, by jointly training on all the datasets. To specify which task the model should perform, we add a task-specific (text) prefix to the original input sequence before feeding it to the model. To handle the im-

balance between tasks in MTL, we use exponential sampling of each task sampling rates. Fine-tuning hyperparameters can be found in Appendix E.

3.1 Experimental Results

We report the results on T5 and PrivaT5 models across small, base, large scales on STL and MTL settings in Table 1. We also report STL results on encoder only models such as BERT (Devlin et al., 2018), LegalBERT (Chalkidis et al., 2020) and PrivBERT (Srinath et al., 2021) which is continually pre-trained on PrivaSeer Corpus.

T5 vs. PrivaT5: STL We observe that PrivaT5-small consistently outperforms T5 across various tasks. The trend is maintained with PrivaT5-Base on most tasks, with the exception of PI-extract and PolicyDetection. Similarly, the large variant follows the same pattern, except for marginal differences on OPP115 and PI-Extract. This underscores the significance of continuous pre-training on domain-specific corpora to achieve superior performance in downstream tasks within that domain. However the degree of improvement varies across tasks. Contrary to expectation, we do not observe any straightforward correlation between size of the dataset and requirement of pre-training as one expects pre-training to benefit in low-data fine-tuning settings. This deviation along with performance decreases on certain configurations prompts a deeper exploration into the intricate dynamics at play during fine-tuning, challenging preconceived notions about the universality of pre-training benefits.

T5 vs. PrivaT5: MTL Except on PolicyDetection in base setting, PrivaT5 outperforms T5 on all tasks in MTL. This clearly demonstrates the utility of domain-specific continued pre-training.

Scaling T5 & PrivaT5: We observe a consistent trend of performance improvement as the scale of parameters increases (from small to base to large) for both T5 and PrivaT5 in both MTL and STL settings. Investigating how the scale of the model translates to the degree of enhancement in these tasks and uncovering the factors influencing these dynamics, presents an interesting direction.

T5 vs. BERT BERT models employed possess 110M parameters, which is double of Small (60M) and half of Base (220M) version of T5. Interestingly, Small version underperforms compared to BERT models, with the Base version catching up, and the Large version attempting comparability across most tasks. Particularly, in tasks involving structured output spaces such as sequence tagging, BERT family models excel, while T5 encounters difficulties in grasping the syntax of complex output spaces. Addressing this challenge necessitates the design of effective decoding mechanisms or better textual transformations of structured output spaces, particularly for information extraction tasks using these generative models. A case in point is PolicyIE-B, where T5-large model despite with 770M parameters underperform compared to BERT family with 110M, highlighting ineffective handling of complex structured output space in generation paradigm, while it is easy to have a token-level classifier for BERT models. In case of PolicyQA, where BERT models can easily be extractive, T5 models generate text similar to the actual answer but aren't inherently extractive. This results in a penalty for T5 models on matching metrics, highlighting the need for nuanced evaluation approaches for different models in various tasks.

STL vs. MTL While MTL underperforms compared to STL in specific configurations, like OPP115 across Small, Base, and Large setups, it shines in contexts such as PolicyIE-A. Contrary to the anticipated positive transfer from MTL, especially in low-data settings through data ensembling, our findings mostly expose negative transfer, aligning with previous studies (Rosenstein et al., 2005; Caruana, 1997). This can be attributed to negative interference between unrelated tasks which dampens task synergies during training, urging a thorough exploration of improved task sampling or grouping strategies (Fifty et al., 2021; Guo et al., 2019; Xu et al., 2019), alongside different optimizations like gradient surgery (Yu et al., 2020) and gradient vaccine (Wang and Tsvetkov, 2021) to counteract negative transfers between tasks.

4 Conclusion

In this study, we introduce PrivaT5, a T5-based transformer model designed for privacy policy text across various scales: small (60M), base (220M), and large (770M). PrivaT5 is obtained by further pre-training T5 on PrivaSeer Corpus of contemporary website privacy policies. We demonstrate that domain-specific pre-trained PrivaT5 models outperform general T5 models on different privacy policy related tasks. Further, we notice that these generative models struggle to handle structured output spaces in case of sequence tagging tasks, indicating a potential avenue for future exploration.

Limitations

While this study offers insights into the effectiveness of PrivaT5 over T5 within privacy policy understanding, we acknowledge its limitations. Our pre-training relies on the PrivaSeer Corpus, which, while comprehensive, may not fully represent the entire spectrum of privacy policy variations. The model's performance could be influenced by potential biases or gaps in the training data. PrivaT5's training and evaluation primarily involve English-language privacy policies. Assessing its performance and generalization capabilities to policies in other languages remains an unexplored area, limiting its applicability in a global context. While our results point to challenges in structured output spaces, particularly in sequence tagging tasks, a deeper investigation into the root causes and potential mitigations is left for future research.

Ethics Statement

PrivaT5 inherits biases present in the training data, potentially perpetuating or amplifying existing biases in privacy policies. Investigating and mitigating these biases is crucial to ensure fair and unbiased model outcomes. The privacy policies used for training may contain sensitive information. While we do not foresee any inherent risks associated, precautionary measures, including data anonymization, are essential to ensure compliance with ethical standards and safeguard against unintended consequences.

References

Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. Policyqa: A reading comprehension dataset for privacy policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 743–749.

Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021*, pages 2165–2176.

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

Duc Bui, Kang G Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated extraction and presentation of data practices in privacy policies. *Proc. Priv. Enhancing Technol.*, 2021(2):88–110.

Carole Cadwalladr and Emma Graham-Harrison. 2018. Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The guardian*, 17(1):22.

Rich Caruana. 1997. Multitask learning (ph. d. thesis).

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Jianfeng Chi, Wasi Uddin Ahmad, Yuan Tian, and Kai-Wei Chang. 2023. PLUE: Language understanding evaluation benchmark for privacy policies in English. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 352–365, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516.

US FTC. Federal trade commission et al. 2012. protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers. *FTC Report*.

Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. 2016. How short is too short? implications of length and framing on the effectiveness of privacy notices. In *Twelfth symposium on usable privacy and security (SOUPS 2016)*, pages 321–340.

Abhijith Athreya Mysore Gopinath, Vinayshekhar Bannihatti Kumar, Shomir Wilson, and Norman Sadeh. 2020. Automatic section title generation to improve the readability of privacy policies. *USENIX SOUPS*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2019. Autosem: Automatic task selection and mixing in multi-task learning. In *Proceedings of NAACL-HLT*, pages 3520–3531.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548.

Mitra Bokaie Hosseini, KC Pragyan, Irwin Reyes, and Serge Egelman. 2020. Identifying and classifying third-party entities in natural language privacy policies. In *Proceedings of the Second Workshop on Privacy in NLP*, pages 18–27.

Moniba Keymanesh, Micha Elsner, and Srinivasan Sarthasarathy. 2020. Toward domain-guided controllable summarization of privacy policies. In *NLLP@ KDD*, pages 18–24.

Vinayshekhar Bannihatti Kumar, Kasturi Bhattacharjee, and Rashmi Gangadharaiah. 2022. Towards cross-domain transferability of text generation models for legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 111–118.

Vinayshekhar Bannihatti Kumar, Abhilasha Ravichander, Peter Story, and Norman Sadeh. 2019. Quantifying the effect of in-domain distributed word representations: A study of privacy policies. In *AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies*.

Martha K Landesberg, Toby Milgrom Levin, Caroline G Curtin, and Ori Lev. 1998. Privacy online: A report to congress. *NASA*, (19990008264).

T Le, T Norton, Y Tian, K Chang, et al. 2021. Intent classification and slot filling for privacy policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Logan Lebanoff and Fei Liu. 2018. Automatic detection of vague words and sentences in privacy policies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3508–3517.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Aaron K Massey, Jacob Eisenstein, Annie I Antón, and Peter P Swire. 2013. Automated text mining for requirements analysis of policy documents. In *2013 21st IEEE International Requirements Engineering Conference (RE)*, pages 4–13. IEEE.

Gabriele Meiselwitz. 2013. Readability assessment of policies and procedures of social networking sites. In *Online Communities and Social Computing: 5th International conference, OCSC 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013. Proceedings 5*, pages 67–75. Springer.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Karthik Raman, Iftekhar Naim, Jiecao Chen, Kazuma Hashimoto, Kiran Yalasangi, and Krishna Srinivasan. 2022. Transforming sequence tagging into a seq2seq task. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11856–11874.

Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah A Smith. 2014. Unsupervised alignment of privacy policies using hidden markov models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 605–610.

Abhilasha Ravichander. 2019. Question answering for privacy policies: Combining computational and legal. In *Proceedings of the 2019 Conference on Empirical Methods in Natural*, pages 4947–4958. Association for Computational Linguistics.

Abhilasha Ravichander, Alan W Black, Thomas Norton, Shomir Wilson, and Norman Sadeh. 2021. Breaking down walls of text: How can nlp benefit consumer privacy? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1.

Joel R Reidenberg, Jaspreet Bhatia, Travis D Breaux, and Thomas B Norton. 2016. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, 45(S2):S163–S190.

Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. 2005. To transfer or not to transfer. In *NIPS 2005 workshop on transfer learning*, volume 898.

Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonald, Joel R Reidenberg, Noah A Smith, Fei Liu, N Cameron Russell, Florian Schaub, et al. 2013. The usable privacy

policy project. In *Technical report, Technical Report, CMU-ISR-13-119*. Carnegie Mellon University.

Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2774–2779.

Atreya Shankar, Andreas Waldis, Christof Bless, Maria Andueza Rodriguez, and Luca Mazzola. 2023. Privacyglue: A benchmark dataset for general language understanding in privacy policies. *Applied Sciences*, 13(6):3701.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Zirui Wang and Yulia Tsvetkov. 2021. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Yichong Xu, Xiaodong Liu, Yelong Shen, Jingjing Liu, and Jianfeng Gao. 2019. Multi-task learning with sample re-weighting for machine reading comprehension. In *Proceedings of NAACL-HLT*, pages 2644–2655.

Task	Train	Dev	Test	#Labels
OPP115	2185	550	697	12
PI-Extract	2579	456	1029	3/3/3/3
Pol.Detection	773	137	391	2
Pol.IE-A	4109	100	1041	5
Pol.IE-B	4109	100	1041	29/9
Priv.QA	17056	3809	4152	-
Pol.QA	157420	27780	62150	2
Pol.Summ	20000	2000	2000	-

Table 2: Statistics of privacy related downstream tasks. PI-Extract and PolicyIE-B consist of four and two subtasks respectively.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836.

Razieh Nokhbeh Zaeem, Rachel L German, and K Suzanne Barber. 2018. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology (TOIT)*, 18(4):1–18.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66.

A Statistics of Downstream Tasks

Table 2 displays dataset splits and number of labels in each of the downstream tasks.

B Examples from Downstream Tasks

Table 3 displays illustrative examples from each of the downstream task, along with each task label space.

C Text-to-text transformation of downstream tasks

Table 4 provide text-to-text transformation of representative examples from each of the downstream task provided in Tab. 3.

D Pre-training Hyperparameters

For all of our pre-trained models, we use a learning rate of 0.001, linear warmup of 2k steps, inverse square root learning rate decay and a maximum sequence length of 512. We employ a batch size of 32, 16 and 8 for small, base and large models respectively and is optimized end-to-end using Adafactor optimizer (Shazeer and Stern, 2018) with

OPP-115	*Secure Online Ordering For your security, we only store your credit card information if you choose to set up an authorized account with one of our Sites. In that case, it is stored on a secure computer in an encrypted format. If you do not set up an account, you will have to enter your credit card information each time you order. We understand that this may be a little inconvenient for you, but some customers appreciate the added security.*
	Labels: Data Retention, Data Security, Do Not Track, First Party Collection/Use, International and Specific Audiences Introductory/Generic, Policy Change, Practice not covered, Privacy contact information, Third Party Sharing/Collection, User Access, Edit and Deletion, User Choice/Control Output: Data Security; User Choice/Control; First Party Collection/Use
PI-Extract	*We may collect and share your IP address but not your email address with our business partners*
	Subtask-I Labels: {B,I}-COLLECT, O Output: O O O O O B-COLLECT I-COLLECT I-COLLECT O O O O O O O O O O Subtask-II Labels: {B,I}-NOT_COLLECT, O Output: O O O O O O O O O O B-NOT_COLLECT I-NOT_COLLECT I-NOT_COLLECT O O O O O Subtask-III Labels: {B,I}-NOT_SHARE, O Output: O O O O O O O O O O B-NOT_SHARE I-NOT_SHARE I-NOT_SHARE O O O O O Subtask-IV Labels: {B,I}-SHARE, O Output: O O O O O B-SHARE I-SHARE I-SHARE O O O O O O O O O O
PolicyDetection	*This website uses Google Analytics, a web analytics service provided by Google, Inc. ("Google"). Google Analytics uses "cookies", which are text. .*
	Labels: Not a Policy, Policy Output: Not a Policy
PolicyIE-A	*CMS websites keep data collected long enough to achieve the specified objective for which they were collected*
	Labels: Data Collection/Usage, Data Security/Protection, Data Sharing/Disclosure, Data Storage/Retention, OtherOutput: Data Storage/retention Output: Data Storage/Retention
PolicyIE-B	*We may also use or display your username and icon or profile photo on marketing purpose or press releases*
	Subtask-I Labels: {B,I}-data-protector, {B,I}-data-protected, {B,I}-data-collector, {B,I}-data-collected, {B,I}-data-receiver, {B,I}-data-retained, {B,I}-data-holder, {B,I}-data-provider, {B,I}-data-sharer, {B,I}-data-shared, {B,I}-storage-place, {B,I}-retention-period, {B,I}-protect-against, {B,I}-action, O Output: B-data-collector O O B-action O O B-data-provider B-data-collected O B-data-collected I-data-collected I-data-collected I-data-collected O O O O O Subtask-II Labels: {B,I}-purpose, {B,I}-polarity, {B,I}-method, {B,I}-condition, O Output: O O O O O O O O O O O O O B-purpose I-purpose I-purpose I-purpose I-purpose

PrivacyQA	*Context : We may collect and use information about your location (such as your country) or infer your approximate location based on your IP address in order to provide you with tailored educational experiences for your region, but we don't collect the precise geolocation of you or your device.* *Question: Does the app track my location?*
	Labels: Relevant, Irrelevant Answer: Relevant
PolicyQA	*Context: Illini Media never shares personally identifiable information provided to us online in ways unrelated to the ones described above without allowing you to opt out or otherwise prohibit such unrelated uses. Google or any ad server may use information (not including your name, address, email address, or telephone number) about your visits to this and other websites in order to provide advertisements about goods and services of interest to you.* *Question: Do you share my data with others? If yes, what is the type of data?*
	Answer: information (not including your name, address, email address or telephone number)
PolicySumm	*You have the right to lodge a complaint with your local data protection supervisory authority, which is the Information Commissioner's Office in the UK.*
	Summary: Right to Complain

Table 3: Illustrative examples of each downstream task

a corrupted token ratio of 15% with the mean noise span length of 3. Pre-training is carried out using Google Cloud TPU with 8 cores (v3.8) from TPU Research Cloud (TRC).[3]

E Fine-tuning Hyperparameters

Each model is trained for 50 epochs, with early stopping and is optimized using Adafactor. We varied learning rate across {1e-3, 5e-4, 3e-4, 1e-4} to identify the optimal rate. Task-specific evaluation metrics are employed for best model selection, with macro-F1 scores for all the tasks except PolicyQA which relied on Exact Match scores. We employ a batch size of 32, 16 and 8 for small, base, and large respectively. All the experiments are carried out on TPU v3-8 device with maximal input sequence length of 512 and truncating lnger sequences beyond. For MTL, we use exponential sampling for data ensemble with $\alpha = 0.01$.

[3]https://sites.research.google/trc

Task Name	Input	Output
OPP-115	*OPP 115 Sentence: Secure Online Ordering For your security, we only store your credit card information if you choose to set up an authorized account with one of our Sites. In that case, it is stored on a secure computer in an encrypted format. If you do not set up an account, you will have to enter your credit card information each time you order. We understand that this may be a little inconvenient for you, but some customers appreciate the added security.*	Data Security; User Choice/Control; First Party Collection/Use
PI-Extract	*PI Extract sentence: <extra_id_0>We <extra_id_1>may <extra_id_2>collect <extra_id_3>and <extra_id_4>share <extra_id_5>your <extra_id_6>IP <extra_id_7>address <extra_id_8>but <extra_id_9>not <extra_id_10>your <extra_id_11>email <extra_id_12>address <extra_id_13>with <extra_id_14>our <extra_id_15>business <extra_id_16>partners*	<extra_id_5>B-COLLECT B-SHARE <extra_id_6>I-COLLECT I-SHARE <extra_id_7>I-COLLECT I-SHARE <extra_id_10>B-NOT_COLLECT B-NOT_SHARE <extra_id_11>I-NOT_COLLECT I-NOT_SHARE <extra_id_12>I-NOT_COLLECT I-NOT_SHARE
PolicyDetection	*Policy Detection : This website uses Google Analytics, a web analytics service provided by Google, Inc. ("Google"). Google Analytics uses "cookies", which are text. .*	Not a Policy
PolicyIE-A	*Policy IE A : CMS websites keep data collected long enough to achieve the specified objective for which they were collected*	Data Storage/Retention
PolicyIE-B	*Policy IE B : <extra_id_0>We <extra_id_1>may <extra_id_2>also <extra_id_3>use <extra_id_4>or <extra_id_5>display <extra_id_6>your <extra_id_7>username <extra_id_8>and <extra_id_9>icon <extra_id_10>or <extra_id_11>profile <extra_id_12>photo <extra_id_13>on <extra_id_14>marketing <extra_id_15>purpose <extra_id_16>or <extra_id_17>press <extra_id_18>releases*	<extra_id_0>B-data-collector <extra_id_3>B-action <extra_id_6>B-data-provider <extra_id_7>B-data-collected <extra_id_9>B-data-collected <extra_id_10>I-data-collected <extra_id_11>I-data-collected <extra_id_12>I-data-collected <extra_id_14>B-purpose <extra_id_15>I-purpose <extra_id_16>I-purpose <extra_id_17>I-purpose <extra_id_18>I-purpose
PrivacyQA	*Privacy QA question: Does the app track my location? Context : We may collect and use information about your location (such as your country) or infer your approximate location based on your IP address in order to provide you with tailored educational experiences for your region, but we don't collect the precise geolocation of you or your device.*	Relevant

PolicyQA	*Policy QA question: Do you share my data with others? If yes, what is the type of data? Context: Illini Media never shares personally identifiable information provided to us online in ways unrelated to the ones described above without allowing you to opt out or otherwise prohibit such unrelated uses. Google or any ad server may use information (not including your name, address, email address, or telephone number) about your visits to this and other websites in order to provide advertisements about goods and services of interest to you.*	information (not including your name, address, email address or telephone number)
PolicySumm	*Title Generation : You have the right to lodge a complaint with your local data protection supervisory authority, which is the Information Commissioner's Office in the UK.*	Right to Complain

Table 4: Text-to-text transformation of illustrative examples for downstream tasks in Tab. 3.