Improving Privacy Language Domain NLU benchmarks for LLMs

Aravind Narasimhan
4/14/2024
University of California, Berkeley

Instructor: Peter Grabowski

Abstract

Privacy Policies across digital applications are hard to digest for consumers due to their complexity and legalese. PLUE (Privacy Policy Language Understanding Evaluation)[1] is a new benchmark meant to enable researchers to evaluate the effectiveness of their NLU techniques in the Privacy domain across tasks including text classification, question answering, semantic parsing and named-entity recognition. As part of this project, I contribute to the PLUE effort by applying the various NLU benchmark tasks against PrivBERT [3], a pre-trained Privacy focused LLM that uses the PrivaSEER[3] corpus and training approach with a much larger and more exhaustive privacy corpus, to show that PrivBERT achieves SOTA PLUE benchmark scores[1] across various NLU domains include Policy Question-Answering, Privacy Question-Answering, Named Entity Recognition and Text Classification.

In addition, I attempt to improve the baseline PLUE benchmarks established for the Privacy Question-Answering [6] task by using a novel approach (in the privacy space) to improve the baseline corpus for readability and understand through sentence rewriting using a current auto-regressive LLM (Meta's LLAMA2 7B) [7]. With Privacy QA corpus rewriting augmentation I was able to achieve 0.6 – 0.7 improvement on the previous privacy QA baseline for F1 score achieved in PLUE and 184 for unanswerable questions with the finetuned PrivBERT baseline.

Introduction

Since 2013, a goal of the Usable Privacy Project [2] has been to *"simplify the extraction, explanation and understanding of privacy policies in an automated manner"*. To this end, much NLU research has been done, along with technologies and dataset that have been developed, to better understand and interpret privacy policies. Various researchers [4],[5],[6],[8],[9],[10],[11] have developed standardized privacy corpuses and adapted various NLU tasks from traditional Natural language processing domains for more specific application to the language complexity of privacy policies. These tasks seek to disentangle privacy policy language with respect to their syntax and semantics. The privacy language specific tasks that have been explored in previous research and that make up the PLUE benchmark include:

- Mult-label text classification (MC)
- Question Answering (QA)
- Name Entity Recognition (NER)
- Semantic Parsing (SP)
  - SP-Intent Classification (IC)
  - SP-Slot-filling (SF)

PLUE: Language Understanding Evaluation Benchmark for Privacy Policies in English

PLUE is a standardized benchmark, dataset and training parameters that allow new researchers in the privacy space to continue to make progress on the effort to improve privacy policy understand and have an objective measure of evaluation on the progress. The PLUE dataset consists of datasets for both pre-training and finetuning privacy models (PLMs) for future evaluation and further benchmarking.

The PLUE dataset is comprised of 64K Mobile Application Privacy policy documents from MAPS[4] and 130K Website privacy policies from the Princeton-Leuven Longitudinal Corpus of Privacy Policies[5] for a total corpus of 332M words that can be used for continuous pretraining of BERT, RoBERTa and other Masked Language Model base LLMs. In addition, the dataset for the benchmark encompass highly cited privacy datasets including **OPP-115** [11] and **APP-**

1. While I was able to surpass the PLUE benchmarks in my GCP AI/ML training environment, I was not able to reproduce the specific metrics that were described in the PLUE paper which were higher than what I was able to produce with the same code and training data in my environment. I am in touch with the principal researcher to understand why this might be.

350 [4] for text classification**; PrivacyQA** [6] and **PolicyQA** [9] for Question Answering; **PolicyIE** [8] for semantic parsing; and **PI-Extract** [12] for Named Entity recognition tasks.

The baseline PLUE benchmarks from [1] are summarized in charts below.

Models	|Model|	OPP-115 F1	APP-350 F1	PrivacyQA P / R / F1	PolicyQA F1 / EM	PI-Extract F1
Human	-	-	-	68.8 / 69.0 / 68.9	-	-
$BERT_{BASE}$	110M	75.3	59.6	44.6 / 35.9 / 36.3	55.1 / 27.7	63.7 / 54.6
$Electra_{BASE}$	110M	74.0	49.3	42.7 / 36.0 / 36.1	57.5 / 29.9	69.4 / 57.8
$SpanBERT_{BASE}$	110M	62.8	32.8	24.8 / 24.8 / 24.8	55.2 / 27.8	66.9 / 41.0
$RoBERTa_{BASE}$	124M	79.0	67.1	43.6 / 36.4 / 36.7	56.6 / 29.4	70.7 / 56.8
$PP\text{-}BERT_{BASE}$	110M	78.0	62.8	44.8 / 36.9 / 37.7	58.3 / 30.0	70.5 / 55.3
$PP\text{-}Electra_{BASE}$	110M	73.1	57.1	48.3 / 38.8 / 39.3	58.0 / 30.0	70.3 / 61.2
$PP\text{-}SpanBERT_{BASE}$	110M	78.1	61.9	43.4 / 36.4 / 36.8	55.8 / 27.5	65.5 / 50.8
$PP\text{-}RoBERTa_{BASE}$	124M	**80.2**	69.5	**49.8 / 40.1 / 40.9**	57.8 / 30.3	**71.2** / 61.3
$LEGAL\text{-}BERT_{BASE}$	110M	76.0	57.4	45.6 / 37.6 / 38.2	55.1 / 27.7	69.1 / 51.1
$BERT_{LARGE}$	340M	79.3	71.2	43.8 / 35.4 / 36.1	56.6 / 28.7	68.1 / 54.8
$Electra_{LARGE}$	340M	78.7	41.5	46.6 / 42.1 / 40.5	**60.7 / 33.2**	70.1 / 59.5
$SpanBERT_{LARGE}$	340M	79.4	66.0	45.2 / 36.5 / 37.3	58.2 / 30.8	68.2 / 50.8
$RoBERTa_{LARGE}$	355M	79.9	**72.4**	47.6 / 41.4 / 40.6	59.8 / 32.5	70.9 / **62.8**

Table 2: Performance comparison of pre-trained models on text classification, question answering, and named entity recognition tasks. We fine-tune all the models three times with different seeds and report average performances. Human performances are reported from the respective works.

Models	|Model|	Intent Classification F1	Slot Filling Type-I Slots F1	Slot Filling Type-I Slots EM	Slot Filling Type-II Slots F1	Slot Filling Type-II Slots EM
Human	-	96.5	84.3	56.6	62.3	55.6
$BERT_{BASE}$	110M	73.7	55.2	19.7	34.7	29.8
$Electra_{BASE}$	110M	73.7	56.4	22.8	36.5	30.7
$SpanBERT_{BASE}$	110M	71.9	44.0	10.8	29.7	17.5
$RoBERTa_{BASE}$	110M	74.5	56.8	22.0	39.2	32.0
$PP\text{-}BERT_{BASE}$	110M	76.9	56.7	22.8	38.7	32.5
$PP\text{-}Electra_{BASE}$	110M	77.1	58.2	**24.1**	37.8	**32.9**
$PP\text{-}SpanBERT_{BASE}$	110M	75.0	54.1	19.8	33.6	26.7
$PP\text{-}RoBERTa_{BASE}$	110M	**78.1**	58.0	22.4	40.1	32.4
$LEGAL\text{-}BERT_{BASE}$	110M	72.6	53.8	19.5	36.1	29.7
$BERT_{LARGE}$	340M	75.5	56.8	23.0	38.4	32.2
$Electra_{LARGE}$	340M	75.6	57.9	24.0	39.6	32.4
$SpanBERT_{LARGE}$	340M	73.8	45.5	9.5	38.8	29.8
$RoBERTa_{LARGE}$	355M	77.6	**58.4**	22.9	**41.4**	32.7

Table 3: Performance comparison of pre-trained models on intent classification and slot filling tasks (PolicyIE). We fine-tune all the models three times with different seeds and report average performances. Human performances are reported from the respective works.

Figure 1: PLUE Benchmarks for various NLP Tasks against general and domain-specific NLPs. The PP-NLPs are Privacy domain specialized NLPs trained and fine-tuned as part of the PLUE Project. Source: PLUE Paper [1]

PP-BERT and PP-RoBERTa

PP-BERT and PP-RoBERTa are Large Language Models (LLMs) continuously pretrained on the PLUE pretraining corpus. They are domain specialized Privacy Language Models (PLMs) that aim to provide higher performance on the NLP tasks in the privacy domain. Since the model itself and the model weights were not published or made available, as part of this project I will be pretraining baseline BERT and RoBERTa to get a baseline PLM for benchmark comparisons[2].

PrivaSEER and PrivBERT

The PrivaSEER corpus of Web Privacy Policies was introduced in 2021 by [3] with the goal of adding to the privacy corpora available for researchers to continue their work on Privacy Language understanding. This corpus comprises of ~1M English language website privacy policies through thorough and vetted document collection process that include language detection, duplicate removal, document classification and content extraction. Notably this corpus of privacy policies is ~10x larger than the next largest public corpus of such documents. As such an PLM trained on this corpus could have the benefit of additional domain data and content to be more effective in understanding privacy language tasks. PrivBERT is exactly such an PLM also introduced and release with [3] in 2021. It is a RoBERTa Base LLM that has been further pre-trained on the PrivaSEER corpus. Thankfully along with the corpus, and the training hyperparameters, the authors also made the actual model and weights available for consumption on HuggingFace.

Motivation & Hypothesis

Opportunity #1: *Playing around with PLUE*

With a new benchmark (PLUE) available, it would be beneficial for the field to have different projects and domain related efforts to start using the Benchmark. That can help us collectively evaluate new techniques and review advancements in the NLP space, as applied to the Privacy domain. As part of my literature review however, I did not see any active privacy project using the PLUE benchmarks. This project breaks ground and starts the process of using PLUE and hopefully will incentivize and motivate other teams to follow suit. In addition, PrivBERT was a great target to for the benchmark efforts as it is a purely privacy domain language model with a similar architecture to PP-BERT and PP-RoBERTa. The main difference between them being the data that PrivBERT has been trained on, which is more expansive and broader and carefully developed as part of the PrivaSeer project.

Hypothesis #1: PrivBERT should achieve better PLUE benchmark metrics across privacy language understanding task than PP-BERT or PP-RoBERTa.

2. Given the tight nature of the project and resource availability, I have only focused on BERT and RoBERTa for additional PLM work. It will be a future effort for someone else to continue review of Electra, SPANBert, and LegalBERT.

Opportunity #2: *Improve Benchmarks on the Privacy Question Answering Task*

Given the privacy policies are hard to understand, one tool that could help with comprehension is the ability to extract correct and relevant answers to questions people may have about a privacy policy. In [6] the authors introduce PrivacyQA, a corpus 1750 questions about privacy policies and 3500 expert annotations about the relevancy of each sentence in the policy to answering the question being asked. QA is hard NLU problem in the privacy space both for Humans and machines. In the 2019 Privacy QA paper [6], Humans achieve an F1 score of only 68.9%. The SOTA ML model (BERT based) used in the paper achieves an even lower F1 score of 39.8% (much worse than humans). PLUE with its PP-RoBERTa PLM achieves a small gain of 1.1% point at 40.9% though both Precision and Recall as substantially improved. As such PrivacyQA seems to be ripe area for review with respect to Privacy NLU techniques and benchmark against new NLP tools and models including the large parameter based LLM models to try to achieve a better benchmark.

Hypothesis #2: In [6] the authors talk about "the goal of this effort (I,e. PrivacyQA) is to kickstart the development of question-answering methods for this domain, to address the (unrealistic) expectation that a large population should be reading many policies per day." In this project, I propose: 1.) hypothesis 2.1 – that better trained PLMs (such as PrivBERT) will automatically beat the PLUE and PrivacyQA paper benchmarks for Privacy Question Answering. 2.) hypothesis 2.2 – that "Sentence Rewriting" of the PrivacyQA Corpus using autoregressive LLMs to make the corpus language more accessible to users will improve the Privacy QA benchmarks without changing any of he original relevancy assignment. Thereby "sentence rewriting" should improve the ability of benchmarking LLMs / PLMs to better answer Privacy related questions.

Methodology

Opportunity #1: Playing around with PLUE

The PLUE dataset can be accessed from https://github.com/JFChi/PLUE. It comes will all the scripts and data necessary to pretrain and fine tune the base models and any additional models you bring in. It was published in May 2023 and is built on PyTorch and Huggingface bootstrapping and trainer scripts.

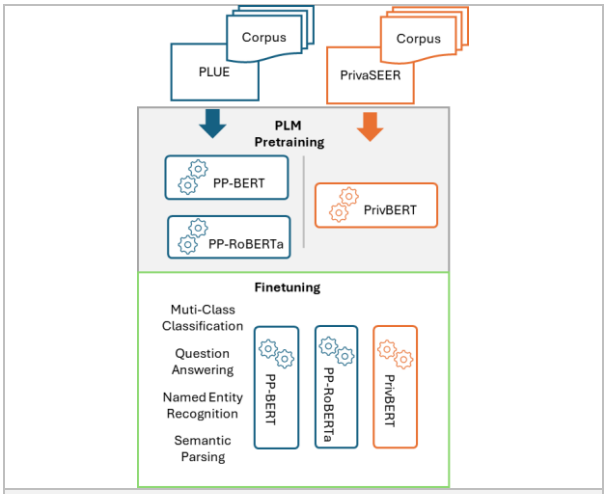

Figure 2: PLUE + PrivBERT benchmarking architecture

PLUE Pretraining

I pre-trained both a BERT (PP-BERT) and RoBERTa (PP-RoBERTa) model using the PLUE default configurations. While PLUE was setup to use 8 concurrent GPUs, I only had access to 1. I also had to reduce the instantaneous batch size and increase the accumulation step to avoid CUDA Out of Memory errors. I trained PP-BERT using for 78hrs, on a P100 GPU 16GB RAM, on the pretraining corpus of PLUE. For PP-RoBERTa, I pretrained for 239hrs.

Table 1: Pretraining of PLMs for PLUE Benchmarks

PP-BERT	**PP-RoBERTa**
Num examples = 828595 Num Epochs = 4 Instantaneous batch = 8 Total train batch size = 32 Gradient Accumulation steps = 4 Total optimization steps = 100000 **Hours: 78 (3.25 days)**	Num examples = 838959 Num Epochs = 4 Instantaneous batch = 8 Total train batch size = 256 Gradient Accumulation steps = 32 Total optimization steps = 12500 **Hours: 239 (1.4 wks)**
Models access on HF https://hugging-face.co/arvyz/plue-ppbert	**Model access on HF** https://hugging-face.co/arvyz/plue-pproberta

PLM Fine-tuning

For PLM finetuning, I used the default settings provided in the PLUE framework for all the various privacy task. Each task has its own adapted scripts and data for finetuning training and model evaluation. It is interesting to note that all fine-tuning tasks

run 3 times (i.e. three models are fine-tuned for each task per each run) and the benchmarks are an average of the 3 runs, of the various associated metrics averaged across three model fine-tuning and predict runs. As it was not easy to review metrics across the runs, I created a helper script, compute_metrics.py to consolidate and report metrics for the various tasks. Most tasks took a few hours to run, but the QA tasks took 9 hours across 3 runs to get complete metrics.

All finetuning was done on 4 x Tesla T4 with 16B GPU RAM and 4xN1 CPU with 15GB cumulative RAM. Across all the pre-training runs and fine-tuning iterations, I used up 350GB of hard drive space!

Opportunity #2: Improve Benchmarks on the Privacy Question Answering Task

To test hypothesis 2.1 I just finetune PrivBERT on the PLUE PrivacyQA task and measure the metrics reported below in the results section.

Testing Hypothesis 2.2 was a bit more of an involved exercise with 5 steps depicted in figure 3. Each step is described in detail below.

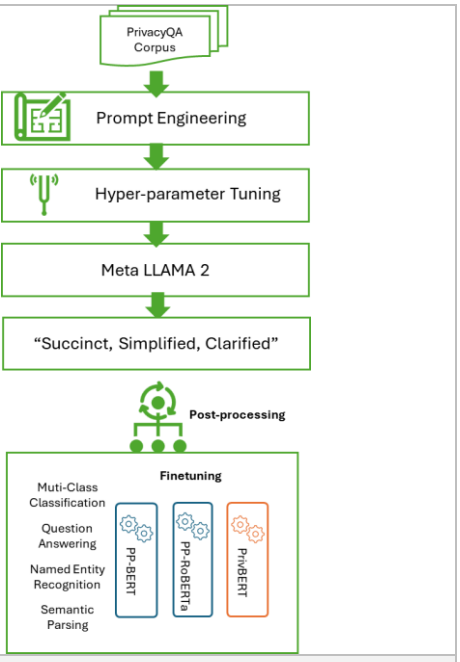

Figure 3: PrivacyQA “Sentence Rewriting” Process

of each task. For example, for the multi-class text classification, F1 scores are computed and

Prompt Engineering

To simplify each of the sentences of the privacy policy, we will use a Large Language Model (LLM) to generate the ouput. The model we will be using is Meta's LLAMA2 7B model that has been fine tuned for "Chat" contexts to be able to take a chat prompt and produce a response.

Prompt Attempts

LLMs are sensitive to prompts for generating appropriate text. So, it is important to test out various prompts to drive the right type of response from the LLM.

The following prompts were tested. The final prompt chosen in bolded:

1. rephrase and simplify this sentence for clarity for an average user:
2. Rephrase and simplify the following sentence for clarity for an average user and do not add any qualifying statements to the response.
3. **Rephrase and simplify the 'SENTENCE' for clarity for an average user and do not add any qualifying statements to the response. SENTENCE:**
4. Rephrase and simplify for clarity for an average user and do not add any qualifying statements to the response. SENTENCE:
5. Rephrase, simplify, and clarify for an average user; do not add any qualifying statements to the response. SENTENCE:

A few additional other prompts were tested, but they did not work at all. So, I have excluded it from the data.

Prompt Selection Process

Sample "rewritten" data was generated for all the above prompts and the final prompt for the QA experiment was selected based on visual inspection of results to ensure the generated text reasonably and faithfully represented the underlying intent of the sentence. ***NOTE:*** This was not done by a privacy expert, but by me in consultation with another layperson. I had to balance between some great simplifications, but some really awkward hallucinations See notebooks for an example.

Hyper Parameter Tuning

To decide the best hyper-parameters for Sentence generation for *simplifying, clarifying, and rephrasing the original sentence succinctly* I tried out multiple variations of hyperparameters that can be used to control the language generation process. These include:

- temperature - Def, 0.25, 0.35, 0.4, 0.5, 0.6
- top_p - Default, 0.8, 0.6, 0.7
- top_k - 5, 7, 10
- num_beams - Default, 2, 3
- early_stopping - True, False

Various combinations of these values were tried with a sample of the training sentence set and manually evaluated for characteristics that would be interesting for the experiment / finetuning process.

Multiple Rewritten Sentence Datasets

I decided to try creating multiple rewritten sentence datasets to test the hypothesis further. The first set of rewriting parameters took a conservative approach to the rewriting rules. The second set of parameters was much more liberal. The goal being to see if "creativity" of vernacular of the LLM would improve QA answerability.

Chosen Values for Sentence rewriting (Version 1 - Conservative) – qa_data_aug.ipynb

- temperature=0.25,
- top_p=0.8,
- top_k=5,
- num_beams=2,
- early_stopping=True,

Chosen Values for Sentence rewriting (Version 2 - Liberal) - qa_data_aug_looser.ipynb

- temperature=0.6,
- top_p=0.7,
- top_k=10,
- num_beams=3,
- early_stopping=False,

LLM Selection

I used the open-source **Meta LLama2 with 7B parameters** to do my sentence rewriting task. This model was chosen because it has performed well in LLM tasks (per published benchmarks by Meta), is easy to load and work with on resource constrained environments and is freely available to researchers to use in their work.

We will specifically use the Huggingface implementation that has been pre-trained on Chat so that we can provide prompts for text generation purposes.

https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

Sentence Post-Processing

The rewritten output sentence, upon inspection, has multiple types of labels that precede sentences like:

- "REPHRASED SENTENCE:",
- "ANALYSIS:",
- "CLEAR AND CLEAR SENTENCE:",
- etc.

We want to extract the sentences without these labels. I do this with Regular expressions (regex).

Results & Analysis

Benchmarking PrivBERT on PLUE

PLUE includes six NLU tasks: Mult-label classification (MC), Question Answering (QA), Intent Classification (IC), Name Entity Recognition (NER) and Slot-filling (SF). For benchmarking PrivBERT, I focus on the multiclass classification as a baseline. I also evaluate the Question Answering, Intent Classification and NER tasks. Results are summarized in the Appendix.

Generally, PrivBERT, trained or a larger corpus of Privacy data outperforms all other models on most Privacy Language task. Some basic tasks like classification see PrivBERT and PP-RoBERTa running neck & neck. PrivBERT is especially strong in NER with the Share task and with PolicyQA where it achieves more that 2% points higher than the closest competitor.

On Privacy QA tasks, PrivBERT continues to exhibit strong performance on Model F1 and unanswerable questions right out the gate. However with, Liberal Sentence rewriting, PrivBERT and the novel sentence rewriting methods adopted in the project,

PrivBERT achieves SOTA Model-F1 score for Privacy the Question Answering task.

Conclusion

As part of the PLUE benchmark, researchers can "continuously pretrain" and further fine-tune BERT and RoBERTa based LLMs to create privacy domain specific LLMs, PP-BERT and PP-RoBERTa respectively.

In this project, I pre-train, fine-tune and apply PP-BERT, PP-RoBERTa on the PLUE dataset to evaluate the impact of domain specific training on potential improvements in NLU metrics for improved privacy policy language understanding. In addition, I use LLM based sentence rewriting technique to improve Privacy QA benchmarks. I demonstrate that I can achieve a SOTA F1 benchmark for Privacy QA using my novel technique.

My code and data can be found at:
https://github.com/arvyz/privacynlu/

Acknowledgements

I would like to thank Jianfeng Chi, principal researcher on the PLUE project for helping me diagnose and suggest a solution for a CUDA-out of memory error when I was pre-training PP-RoBERTa.

I would also like to thank Mukund Srikanth for getting back to me quickly on my questions regarding the PivaSEER project.

GCP was generous in its $300 new project allocation and in quickly responding to my requests for quota increases for GPUs.

Also, I am not sure anyone ever acknowledges or thanks Huggingface [13], but their transformer, the public hosted environment and their AI libraries are the bomb! None of this would be easily possible for low-budget, low-resource researchers without the amazing environment they have put together.

ChatGPT – These days, I don't know how people can code without ChatGPT or the various Co-Pilots out there. Chat (as it is affectionately call) was immensely helpful in writing my code for both my compute_metrics.py and qa_data_aug.ipynb. It was also immensely helpful in resurfacing my memories of bash syntax, python syntax and arcane options and parameters for pytorch (which I had to learn for this project) pipelines.

References:

1. 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.
2. Sadeh, Norman, et al. "The usable privacy policy project." Technical report, Technical Report, CMU-ISR-13-119. Carnegie Mellon University, 2013.
3. 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, Online. Association for Computational Linguistics.
4. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. Privacy Enhancing Technologies Symposium 2019.
5. Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., & Mayer, J. (2021, April). Privacy policies over time: Curation and analysis of a million-document dataset. In Proceedings of the Web Conference 2021 (WWW '21). ACM. http://dx.doi.org/10.1145/3442381.3450048
6. Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-

IJCNLP), pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.

7. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Fasal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.
8. Wasi Ahmad, Jianfeng Chi, Tu Le, Thomas Norton, Yuan Tian, and Kai-Wei Chang. 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 (Volume 1: Long Papers), pages 4402– 4417, Online. Association for Computational Linguistics.
9. 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, Online. Association for Computational Linguistics
10. 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.
11. Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016a. The creation and analysis of a website privacy policy corpus. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume
12. 1: Long Papers), pages 1330–1340, Berlin, Germany. Association for Computational Linguistics.
13. Duc Bui, Kang G. Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated extraction and presentation of data practices in privacy policies. Proceedings on Privacy Enhancing Technologies, 2021(2):88–110
14. Huggingface - https://huggingface.co/; Implementations of Transformers; BERT, RoBERTa, PrivBERT, LLAMA2

Appendix – Detailed Results

*Multiclass Classification – **app350***

Model	f1 (macro)	f1 (micro)	f1 (weighted)
bert_lr3e-5	0.6630	0.8518	0.8609
policy_roberta_lr3e-5	0.6661	0.8560	0.8649
privbert_lr3e-5	0.6934	0.8632	0.8700

On baseline Mobile App privacy policies PrivBert Outperforms PP-RoBERTa and basic BERT

Multiclass Classification – opp115

Model	f1 (macro)	f1 (micro)	f1 (weighted)
bert_lr3e-5	0.7450	0.7597	0.7572
policy_bert_lr3e-5	0.7633	0.7853	0.7835
policy_roberta_lr3e-5	0.7755	0.7892	0.7886
privbert_lr3e-5	0.7737	0.7872	0.7861
roberta_lr3e-5	0.7662	0.7783	0.7769

On baseline web privacy policies, PP-RoBERTa just barely outperforms PrivBERT. The domain specific PLMs outperform the standard LLMs

Named Entity Recognition – PI-Extract - CollectUse

Model	accuracy	f1	precision	recall
bert_lr5e-5	0.9816	0.6547	0.6359	0.6748
policy_bert_lr5e-5	0.9837	0.6900	0.6677	0.7139
policy_roberta_lr5e-5	0.9849	0.7115	0.6952	0.7286
privbert_lr5e-5	0.9842	0.7111	0.6930	0.7302
roberta_lr5e-5	0.9828	0.7090	0.6871	0.7327

On NER Collect/Use scenario, PP-RoBERTa just barely outperforms PrivBERT. The domain specific PLMs all do relatively well. But vanilla RoBERTa does stand out to be pretty close.

Named Entity Recognition – PI-Extract - Share

Model	accuracy	f1	precision	recall
bert_lr5e-5	0.9880	0.5317	0.5808	0.4912
policy_bert_lr5e-5	0.9877	0.5283	0.5741	0.4896
policy_roberta_lr5e-5	0.9891	0.6036	0.6164	0.5917
privbert_lr5e-5	0.9892	0.6152	0.6289	0.6029
roberta_lr5e-5	0.9877	0.5643	0.6086	0.5263

On NER Share Scenario, PrivBERT outperforms all other models. The RoBERTa variant PLMs perform significantly better than the rest.

Semantic Parsing – Intent Classification

Model	accuracy	f1
bert_lr3e-5	0.8546	0.7410
policy_bert_lr3e-5	0.8723	0.7672
policy_roberta_lr3e-5	0.8700	0.7649
privbert_lr3e-5	0.8745	0.7777
roberta_lr3e-5	0.8553	0.7463

For Intent Classification, PrivBERT outperforms all other models. Thought the distance is not too far with the other models.

Question Answering – Policy QA

Model	exact_match	f1
bert_lr3e-5	27.0873	54.5709
policy_bert_lr3e-5	29.4236	57.0200
policy_roberta_lr3e-5	28.7653	56.3455
privbert_lr3e-5	31.0051	58.4745
roberta_lr3e-5	29.8892	56.7731

For Question Answering with the Policy QA dataset, PrivBERT significantly outperforms all other models.

Question Answering – Privacy QA – Original Corpus

Model	accuracy	f1	ua-pred	ua-true	Model-Precision	Model-Recall	Model-F1	Human-Precision	Human-Recall	Human-F1
bert_lr2e-5	0.9684	0.6032	179.3333	34.0000	44.0733	36.0267	36.4267	68.8100	69.0400	68.9200
policy_bert_lr2e-5	0.9700	0.6220	203.6667	34.0000	46.2567	37.8833	38.8333	68.8100	69.0400	68.9200
policy_roberta_lr2e-5	0.9692	0.6246	194.3333	34.0000	46.3000	37.8000	38.4833	68.8100	69.0400	68.9200
privbert_lr2e-5	0.9687	0.6189	184.0000	34.0000	47.9600	39.0533	39.6700	68.8100	69.0400	68.9200
roberta_lr2e-5	0.9681	0.6035	204.0000	34.0000	42.6267	36.2733	36.3167	68.8100	69.0400	68.9200

For Question Answering with the Privacy QA dataset with the Original Corpus, PrivBERT outperforms all other models and gets the lowest number of unanswered questions.

Question Answering – Privacy QA – Sentence Rewriting Conservative

Model	accuracy	f1	ua-pred	ua-true	Model-Precision	Model-Recall	Model-F1	Human-Precision	Human-Recall	Human-F1
bert_lr2e-5	0.9688	0.5894	193.3333	34.0000	42.7633	34.7000	35.1300	68.8100	69.0400	68.9200
policy_roberta_lr2e-5	0.9703	0.6202	208.0000	34.0000	47.0733	39.3367	39.4100	68.8100	69.0400	68.9200
privbert_lr2e-5	0.9690	0.6154	188.3333	34.0000	47.0800	39.9867	39.6633	68.8100	69.0400	68.9200
roberta_lr2e-5	0.9685	0.5995	201.6667	34.0000	43.5800	37.3700	36.9733	68.8100	69.0400	68.9200

For Question Answering with the Privacy QA dataset with the with conservative sentence rewriting, PrivBERT and other models perform about the same as previously. BERT degrades noticeably. NOTE: However I believe there was one small data error in the training set that could have affected the finetuning performance. This was discovered later, but without enough time to update the models.

Question Answering – Privacy QA – Sentence Rewriting Liberal

Model	accuracy	f1	ua-pred	ua-true	Model-Precision	Model-Recall	Model-F1	Human-Precision	Human-Recall	Human-F1
bert_lr2e-5	0.9682	0.5970	189.3333	34.0000	45.0467	36.0267	36.7233	68.8100	69.0400	68.9200
policy_roberta_lr2e-5	0.9680	0.6159	198.5000	34.0000	46.4750	39.5000	39.2200	68.8100	69.0400	68.9200
privbert_lr2e-5	0.9669	0.6072	186.6667	34.0000	48.5167	41.1667	41.5733	68.8100	69.0400	68.9200
roberta_lr2e-5	0.9679	0.5908	206.0000	34.0000	43.2567	36.5100	37.0433	68.8100	69.0400	68.9200

For Question Answering with the Privacy QA dataset with the with liberal sentence rewriting, PrivBERT and other models outperform the original baseline and achieve SOTA Model F1 scores compared with prior research and benchmarks.