Article

PrivacyGLUE: A Benchmark Dataset for General Language Understanding in Privacy Policies

Atreya Shankar [*,†] , Andreas Waldis [†] , Christof Bless [†] , Maria A. Rodriguez [†] , Luca Mazzola [*,†]

Posted Date: 3 March 2023

doi: 10.20944/preprints202303.0046.v1

Keywords: Privacy Policies; NLP; benchmark; general language understanding; domain specialization and generalization

Article

PrivacyGLUE: A Benchmark Dataset for General Language Understanding in Privacy Policies

Atreya Shankar [1,*,†], Andreas Waldis [1,†], Christof Bless [1,†], Maria A. Rodriguez [1,†] and Luca Mazzola [1,*,†]

[1] Information Systems Research Lab, HSLU - Lucerne University of Applied Sciences and Arts, Suurstoffi 1, CH-6343 Rotkreuz, Switzerland; andreas.waldis@hslu.ch (A.W.); christof.bless@hslu.ch (C.B.); maria.anduezarodriguez@hslu.ch (M.A.R.)

* Correspondence: atreya.shankar@hslu.ch (A.S.); luca.mazzola@hslu.ch (L.M.); Tel.: +41-41-757-30-97 (A.S.); +41-41-757-68-90 (L.M.)

† Each author contributed equally to this work.

Featured Application: We propose the PrivacyGLUE benchmark to compare and contrast NLP models' general language understanding in the privacy language domain. This will help practitioners in selecting understanding models for applications within the privacy language domain.

Abstract: Benchmarks for general language understanding have been rapidly developing in recent years of NLP research, particularly because of their utility in choosing strong-performing models for practical downstream applications. While benchmarks have been proposed in the legal language domain, virtually no such benchmarks exist for privacy policies despite their increasing importance in modern digital life. This could be explained by privacy policies falling under the legal language domain, but we find evidence to the contrary that motivates a separate benchmark for privacy policies. Consequently, we propose PrivacyGLUE as the first comprehensive benchmark of relevant and high-quality privacy tasks for measuring general language understanding in the privacy language domain. Furthermore, we release performances from multiple transformer language models and perform model-pair agreement analysis to detect tasks where models benefited from domain specialization. Our findings show the importance of in-domain pretraining for privacy policies. We believe PrivacyGLUE can accelerate NLP research and improve general language understanding for humans and AI algorithms in the privacy language domain, thus supporting the adoption and acceptance rates of solutions based on it.

Keywords: Privacy Policies; NLP; benchmark; general language understanding; domain specialization and generalization

1. Introduction

Data privacy is evolving into a critical aspect of modern life with the United Nations (UN) describing it as a *human right in the digital age* [1]. Despite its importance, several studies have demonstrated high barriers to the understanding of privacy policies [2] and estimate that an average person would require ∼200 hours annually to read through all privacy policies encountered in their daily life [3]. To address this, studies such as Wilson et al. [4] recommend training Artificial Intelligence (AI) algorithms on appropriate benchmark datasets to assist humans in understanding privacy policies.

In recent years, benchmarks have been gaining popularity in Machine Learning and Natural Language Processing (NLP) communities because of their ability to holistically evaluate model performance over a variety of representative tasks, thus allowing practitioners to compare and contrast different models on multiple tasks relevant for the specific application domain. GLUE [5] and SuperGLUE [6] are examples of popular NLP benchmarks which measure the natural language understanding capabilities of SOTA models. NLP benchmarks are also developing rapidly in language domains, with LexGLUE [7] being an example of a recent benchmark hosting several difficult tasks in the legal language domain. Interestingly, we do not find similar NLP benchmarks in the privacy language domain for privacy policies. While this could be explained by privacy policies falling under the legal language domain due to their formal and jargon-heavy nature, we claim that privacy policies fall under a distinct language domain and cannot be subsumed under any other specialized NLP benchmark such as LexGLUE.

To investigate this claim, we gather documents from Wikipedia [8], European Legislation (EURLEX; Chalkidis et al.[9]) and company privacy policies [10], with each corpus truncated to 2.5M tokens. Next, we feed these documents into BERT and gather contextualized embeddings, which are then projected to 2-dimensional space using UMAP [11]. In Figure 1, we observe that the three domain corpora cluster independently, providing evidence that privacy policies lie in a distinct language domain from both legal and wikipedia documents. With this motivation, we propose PrivacyGLUE as the first comprehensive benchmark for measuring general language understanding in the privacy language domain. Our main contributions are threefold:

1. Composition of seven high-quality and relevant PrivacyGLUE tasks, specifically OPP-115, PI-Extract, Policy-Detection, PolicyIE-A, PolicyIE-B, PolicyQA and PrivacyQA.
2. Benchmark performances of five transformer language models on all aforementioned tasks, specifically BERT, RoBERTa, Legal-BERT, Legal-RoBERTa and PrivBERT.
3. Model agreement analysis to detect PrivacyGLUE task examples where models benefited from domain specialization.

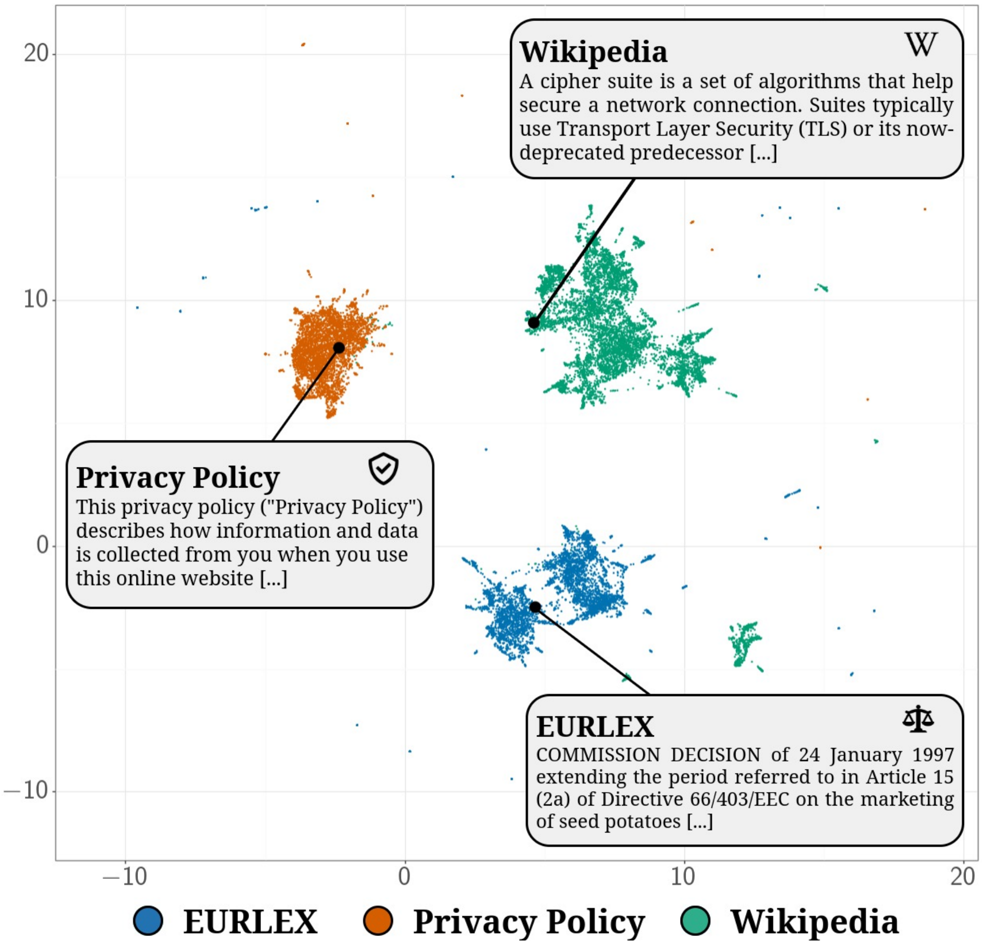

Figure 1. UMAP visualization of BERT embeddings from Wikipedia, European Legislation (EURLEX) and company privacy policy documents with a total of 2.5M tokens per corpus

We release PrivacyGLUE as a fully configurable benchmark suite for straight-forward reproducibility and production of new results in our public GitHub repository[1]. Our findings show that PrivBERT, the only model pretrained on privacy policies, outperforms other models by an average of $2-3\%$ over all PrivacyGLUE tasks, shedding light on the importance of in-domain pretraining for privacy policies. Our model-pair agreement analysis explores specific examples where PrivBERT's privacy-domain pretraining provided both competitive advantage and disadvantage. By benchmarking holistic model performances, we believe PrivacyGLUE can accelerate NLP research into the privacy language domain and ultimately improve general language understanding of privacy policies for both humans and AI algorithms.

2. Related work

NLP benchmarks have been gaining popularity in recent years because of their ability to holistically evaluate model performance over a variety of representative tasks. GLUE [5] and SuperGLUE [6] are examples of benchmarks that evaluate SOTA models on a range of natural language understanding tasks. The GEM benchmark [12] looks beyond text classification and measures performance in Natural Language Generation tasks such as summarization and data-to-text conversion. The XTREME [13] and XTREME-R [14] benchmarks specialize in measuring cross-lingual transfer learning on 40-50 typologically diverse languages and corresponding tasks. Popular NLP benchmarks often host public leaderboards with SOTA scores on supported tasks, thereby encouraging the community to apply new approaches for surpassing top scores.

While the aforementioned benchmarks focus on problem types such as natural language understanding and generation, other benchmarks focus on language domains. The LexGLUE benchmark [7] is an example of a benchmark that evaluates models on tasks from the legal language domain. LexGLUE consists of seven English-language tasks that are representative of the legal language domain and chosen based on size and legal specialization. Chalkidis et al. [7] benchmarked several models such as BERT [15] and Legal-BERT [16], where Legal-BERT has a similar architecture to BERT but was pretrained on diverse legal corpora. A key finding of LexGLUE was that Legal-BERT outperformed other models which were not pretrained on legal corpora. In other words, they found that an in-domain pretrained model outperformed models that were pretrained out-of-domain.

In the privacy language domain, we tend to find isolated datasets from specialized studies. Zimmeck et al. [17], Wilson et al. [4], Bui et al. [18] and Ahmad et al. [19] are examples of studies that introduce annotated corpora for privacy-practice sequence and token classification tasks, while Ravichander et al. [20] and Ahmad et al. [21] release annotated corpora for privacy-practice question answering. Amos et al. [22] is another recent study that released an annotated corpus of privacy policies. As of writing, no comprehensive NLP benchmark exists for general language understanding in privacy policies, making PrivacyGLUE the first consolidated NLP benchmark in the privacy language domain.

3. Datasets and Tasks

The PrivacyGLUE benchmark consists of seven natural language understanding tasks originating from six datasets in the privacy language domain. Summary statistics, detailed label information and representative examples are shown in Table 1, Table A1 (Appendix B) and Table A2 (Appendix C) respectively.

OPP-115

Wilson et al. [4] was the first study to release a large annotated corpus of privacy policies. A total of 115 privacy policies were selected based on their corresponding company's popularity on

[1] Repository will be made public post-acceptance. Anonymous repository: https://anonymous.4open.science/r/f4293357886f671347fa69fae3650543

Google Trends. The selected privacy policies were annotated with 12 data privacy practices on a paragraph-segment level by experts in the privacy domain. As noted by Mousavi Nejad et al. [23], one limitation of Wilson et al. [4] was the lack of publicly released training and test data splits which are essential for machine learning and benchmarking. To address this, Mousavi Nejad et al. [23] released their own training, validation and test data splits for researchers to easily reproduce OPP-115 results. PrivacyGLUE utilizes the "Majority" variant of data splits released by Mousavi Nejad et al. [23] to compose the OPP-115 task. Given an input paragraph segment of a privacy policy, the goal of OPP-115 is to predict one or more data practice categories.

Table 1. Summary statistics of PrivacyGLUE benchmark tasks; † PI-Extract and PolicyIE-B consist of four and two subtasks respectively and the number of BIO token classes per subtask are separated by a forward slash character

Task	Source	Task Type	Train/Dev/Test Instances	# Classes
OPP-115	Wilson et al. [4]	Multi-label sequence classification	2,185/550/697	12
PI-Extract	Bui et al. [18]	Multi-task token classification	2,579/456/1,029	3/3/3/3†
Policy-Detection	Amos et al. [22]	Binary sequence classification	773/137/391	2
PolicyIE-A	Ahmad et al. [19]	Multi-class sequence classification	4,109/100/1,041	5
PolicyIE-B	Ahmad et al. [19]	Multi-task token classification	4,109/100/1,041	29/9†
PolicyQA	Ahmad et al. [21]	Reading comprehension	17,056/3,809/4,152	–
PrivacyQA	Ravichander et al. [20]	Binary sequence classification	157,420/27,780/62,150	2

PI-Extract

Bui et al. [18] focuses on enhanced data practice extraction and presentation to help users better understand privacy policies. As part of their study, they released the PI-Extract dataset consisting of 4.1K sentences (97K tokens) and 2.6K expert-annotated data practices from 30 privacy policies in the OPP-115 dataset. Expert annotations were performed on a token-level for all sentences of selected privacy policies. PI-Extract is broken into four subtasks, where spans of tokens are independently tagged using the BIO scheme commonly used in Named Entity Recognition (NER). Subtasks I, II, III and IV require the classification of token spans for data-related entities that are collected, not collected, not shared and shared respectively. In the interest of diversifying tasks in PrivacyGLUE, we composed PI-Extract as a multi-task token classification problem where all four PI-Extract subtasks are to be jointly learned.

Policy-Detection

Amos et al. [22] developed a crawler for automated collection and curation of privacy policies. An important aspect of their system is the automated classification of documents into privacy policies and non-privacy-policy documents encountered during web crawling. To train such a privacy policy classifier, Amos et al. [22] performed expert annotations of commonly encountered documents during web crawls and classified them into the aforementioned categories. The Policy-Detection dataset was released with a total of 1.3K annotated documents and is utilized in PrivacyGLUE as a binary sequence classification task.

PolicyIE

Inspired by Wilson et al. [4] and Bui et al. [18], Ahmad et al. [19] created PolicyIE, an English corpus composed by 5.3K sentence-level and 11.8K token-level data practice annotations over 31 privacy policies from websites and mobile applications. PolicyIE was designed to be used for machine learning in NLP, to ultimately make data privacy concepts easier for users to understand. We split the PolicyIE corpus into two tasks, namely *PolicyIE-A* and *PolicyIE-B*. Given an input sentence, PolicyIE-A entails multi-class data practice classification while PolicyIE-B entails multi-task token classification over distinct subtasks I and II, which require the classification of token spans for entities that participate in privacy practices and their conditions/purposes respectively. The motivation for composing PolicyIE-B as a multi-task problem is similar to that of PI-Extract.

PolicyQA

Ahmad et al. [21] argue in favour of short-span answers to user questions for long privacy policies. They release PolicyQA, a dataset of 25k reading comprehension examples curated from the OPP-115 corpus from Wilson et al. [4]. Furthermore, they provide 714 human-written questions optimized for a wide range of privacy policies. The final question-answer annotations follow the SQuAD-1.0 format [24], which improves the ease of adaptation into NLP pipelines. We utilize PolicyQA as PrivacyGLUE's reading comprehension task.

PrivacyQA

Similar to Ahmad et al. [21], Ravichander et al. [20] argue in favour of annotated question-answering data for training NLP models to answer user questions about privacy policies. They correspondingly released PrivacyQA, a corpus composed by 1.75K questions and more than 3.5K expert annotated answers. Unlike PolicyQA, PrivacyQA proposes a binary sequence classification task where a question-answer pair is classified as either relevant or irrelevant. Correspondingly, we treat PrivacyQA as a binary sequence classification task in PrivacyGLUE.

4. Experimental setup

The PrivacyGLUE benchmark was tested using the BERT, RoBERTa, Legal-BERT, Legal-RoBERTa and PrivBERT models which are summarized in Table 2. We describe the models used and task-specific approaches, and provide details on our benchmark configuration in Appendix A.

Table 2. Summary of models used in the PrivacyGLUE benchmark; all models used are base-sized variants of BERT/RoBERTa architectures; † BC = BookCorpus, CC-News = CommonCrawl-News, OWT = OpenWebText; ‡ models were initialized with the pretrained RoBERTa model

Model	Source	# Params	Vocab. Size	Pretraining corpora†
BERT	Devlin et al. [15]	110M	30K	Wikipedia, BC (16 GB)
RoBERTa	Liu et al. [25]	125M	50K	Wikipedia, BC, CC-News, OWT (160 GB)
Legal-BERT	Chalkidis et al. [16]	110M	30K	Legislation, Court Cases, Contracts (12 GB)
Legal-RoBERTa‡	Geng et al. [26]	125M	50K	Patents, Court Cases (5 GB)
PrivBERT‡	Srinath et al. [27]	125M	50K	Privacy policies (17 GB)

4.1. Models

BERT

Proposed by Devlin et al. [15], BERT is perhaps the most well-known transformer language model. BERT utilizes the WordPiece tokenizer [28] and is case-insensitive. It is pretrained with the Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks on the Wikipedia and BookCorpus corpora.

RoBERTa

Liu et al. [25] proposed RoBERTa as an improvement to BERT. RoBERTa uses dynamic token masking and eliminates the NSP task during pretraining. Furthermore, it uses a case sensitive byte-level Byte-Pair Encoding [29] tokenizer and is pretrained on larger corpora. Liu et al. [25] reported improved results on various benchmarks using RoBERTa over BERT.

Legal-BERT

Chalkidis et al. [16] proposed Legal-BERT by pretraining BERT from scratch on legal corpora consisting of legislation, court cases and contracts. The sub-word vocabulary of Legal-BERT is learned from scratch using the SentencePiece [30] tokenizer to better support legal terminology. Legal-BERT was the best overall performing model in the LexGLUE benchmark as reported in Chalkidis et al. [7].

Legal-RoBERTa

Inspired by Legal-BERT, Geng et al. [26] proposed Legal-RoBERTa by further pretraining RoBERTa on legal corpora, specifically patents and court cases. Legal-RoBERTa is pretrained on less legal data than Legal-BERT while producing similar results on downstream fine-tuning legal domain tasks.

PrivBERT

Due to the scarcity of large corpora in the privacy domain, Srinath et al. [27] proposed PrivaSeer, a novel corpus of 1M English language website privacy policies crawled from the web. They subsequently proposed PrivBERT by further pretraining RoBERTa on the PrivaSeer corpus.

4.2. *Task-specific approaches*

Given the aforementioned models and tasks, we now describe our task-specific fine-tuning and evaluation approaches. Given an input sequence $s = \{w_1, w_2, \ldots, w_N\}$ consisting of N sequential sub-word tokens, we feed s into a transformer encoder and obtain a contextual representation $\{h_0, h_1, \ldots, h_N\}$ where $h_i \in \mathbb{R}^D$ and D is the output dimensionality of the transformer encoder. Here, h_0 refers to the contextual embedding for the starting token which is `[CLS]` for BERT-derived models and `<s>` for RoBERTa-derived models. For PolicyQA and PrivacyQA, the input sequence s is composed by concatenating the question and context/answer pairs respectively. The concatenated sequences are separated by a separator token, which is `[SEP]` for BERT-derived models and `</s>` for RoBERTa-derived models.

4.2.1. Sequence classification

The h_0 embedding is fed into a class-wise sigmoid classifier (1) and softmax classifier (2) for multi-label and binary/multi-class tasks respectively. The classifier has weights $W \in \mathbb{R}^{D \times C}$ and bias $b \in \mathbb{R}^C$ and is used to predict the probability vector $y \in \mathbb{R}^C$, where C refers to the number of output classes. We fine-tune models end-to-end by minimizing the binary cross-entropy loss and cross-entropy loss for multi-label and binary/multi-class tasks respectively.

$$y = \text{sigmoid}\left(W^\top h_0 + b\right) \tag{1}$$

$$y = \text{softmax}\left(W^\top h_0 + b\right) \tag{2}$$

We report the macro and micro-average F_1 scores for all sequence classification tasks since the former ignores class imbalance while the latter takes it into account.

4.2.2. Multi-task token classification

Each $h_i \in \{h_1, h_2, \ldots, h_N\}$ token embedding is fed into J independent softmax classifiers with weights $W_j \in \mathbb{R}^{D \times C_j}$ and bias $b_j \in \mathbb{R}^{C_j}$ to predict the token probability vector $y_{ij} \in \mathbb{R}^{C_j}$, where C_j refers to the number of output BIO classes per subtask $j \in \{1, 2, \ldots, J\}$. We fine-tune models end-to-end by minimizing the cross-entropy loss across all tokens and subtasks.

$$y_{ij} = \text{softmax}\left(W_j^\top h_i + b_j\right) \tag{3}$$

We report the macro and micro-average F_1 scores for all multi-task token classification tasks by averaging the respective metrics for each subtask. Furthermore, we ignore cases where B or I prefixes are mismatched as long as the main token class is correct.

4.2.3. Reading comprehension

Each $h_i \in \{h_1, h_2, \dots, h_N\}$ token embedding is fed into two independent linear layers with weights $W_j \in \mathbb{R}^D$ and bias $b_j \in \mathbb{R}$ where $j \in \{1, 2\}$. These linear outputs are then concatenated per layer and a softmax function is applied to form a probability vector y_j across all tokens for answer-start and answer-end token probabilities respectively. We fine-tune models end-to-end by minimizing the cross-entropy loss on the gold answer-start and answer-end indices.

$$y_j = \text{softmax}\Big(\big[W_j \cdot h_1 + b_j \quad \dots \quad W_j \cdot h_N + b_j\big]\Big) \tag{4}$$

Similar to SQuAD [24], we report the sample F_1 and exact match accuracy for our reading comprehension task. It is worth noting that Rajpurkar et al. [24] refer to their reported F_1 score as a macro-average, whereas we refer to it as the sample-average as we believe this is a more accurate term.

5. Results

After running the PrivacyGLUE benchmark with 10 random seeds, we collect results on the test-sets of all tasks. Figure 2 shows the respective results in a graphical form while Table A4 in Appendix E shows the numerical results in a tabular form. In terms of absolute metrics, we observe that PrivBERT outperforms other models for all PrivacyGLUE tasks. We apply the Mann-Whitney U-test [31] over random seed metric distributions and find that PrivBERT significantly outperforms other models on six out of seven PrivacyGLUE tasks with $p <= 0.05$, where Policy-Detection was the task where the significance threshold was not met. We utilize the Mann-Whitney U-test because it does not require a normal distribution for test-set metrics, an assumption which has not been extensively validated for deep neural networks [32].

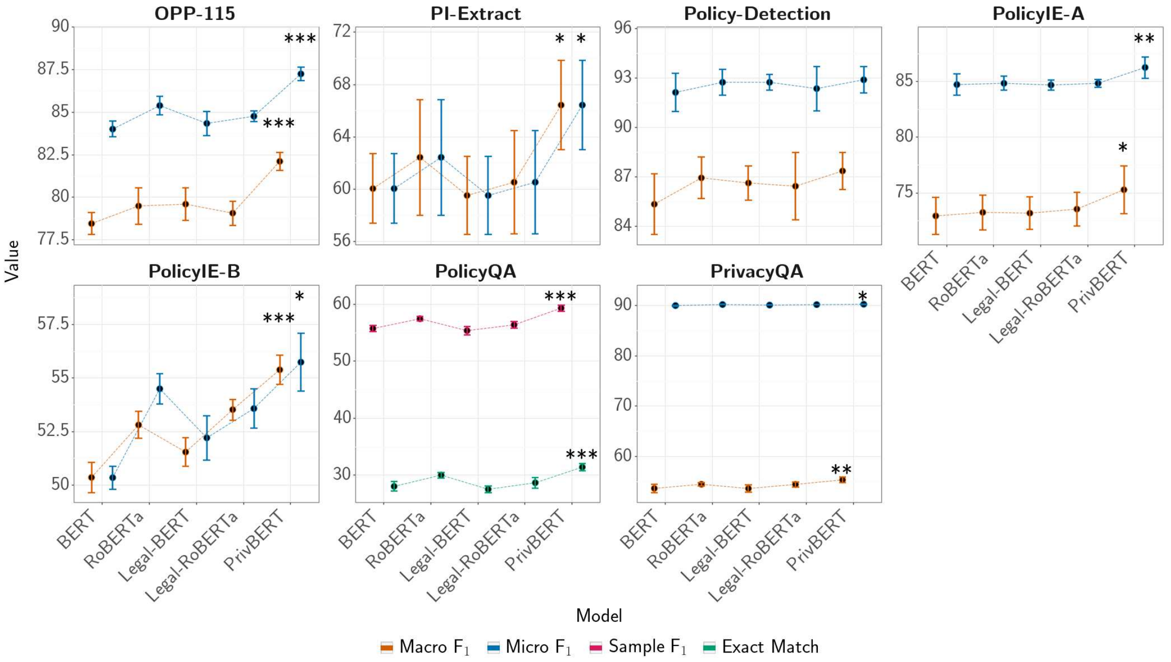

Figure 2. Test-set results of the PrivacyGLUE benchmark where points indicate mean performance and error bars indicate standard deviation over 10 random seeds; *** implies $p <= 0.001$, ** implies $0.001 < p <= 0.01$, * implies $0.01 < p <= 0.05$ given an alternative hypothesis that PrivBERT has a greater performance metric than all other models in a task using the Mann-Whitney U-test

In Figure 2, we observe large differences between the two representative metrics for OPP-115, Policy-Detection, PolicyIE-A, PrivacyQA and PolicyQA. For the first four of the aforementioned tasks, this is because of data imbalance resulting in the micro-average F_1 being significantly higher since it can be skewed by the metric of the majority class. For PolicyQA, this occurs because the EM metric requires exact matches and is therefore much stricter than the sample F_1 metric. Furthermore, we observe an exceptionally large standard deviation on PI-Extract metrics compared to other tasks. This can be attributed to data imbalance between the four subtasks of PI-Extract, with the `NOT_COLLECT` and `NOT_SHARE` subtasks having less than 100 total examples each.

We apply the arithmetic, geometric and harmonic means to aggregated metric means and standard deviations as shown in Table 3. With this, we observe the following general ranking of models from best to worst: PrivBERT, RoBERTa, Legal-RoBERTa, Legal-BERT and BERT. Interestingly, models derived from RoBERTa generally outperformed models derived from BERT. Using the arithmetic mean for simplicity, we observe that PrivBERT outperforms all other models by $2 - 3\%$.

Table 3. Macro-aggregation of means (μ) and standard deviations (σ) per model using the arithmetic mean (A-Mean), geometric mean (G-Mean) and harmonic mean (H-Mean)

Model	A-Mean		G-Mean		H-Mean	
	μ	σ	μ	σ	μ	σ
BERT	67.5	1.1	64.6	0.9	61.1	0.6
RoBERTa	69.0	1.2	66.4	0.7	63.2	0.3
Legal-BERT	67.9	1.1	64.9	0.8	61.2	0.4
Legal-RoBERTa	68.5	1.3	65.7	0.8	62.3	0.4
PrivBERT	**70.8**	1.2	**68.3**	0.8	**65.2**	0.5

6. Discussion

With the PrivacyGLUE benchmark results, we revisit our privacy vs. legal language domain claim from Section 1 and discuss our model-pair agreement analysis for detecting PrivacyGLUE task examples where models benefited from domain specialization.

6.1. Privacy vs. legal language domain

We initially provided evidence from Figure 1 suggesting that the privacy language domain is distinct from the legal language domain. We believe that our PrivacyGLUE results further support this initial claim. If the privacy language domain was subsumed under the legal language domain, we could have observed Legal-RoBERTa and Legal-BERT performing competitively with PrivBERT. Instead, we observed that the legal models underperformed compared to both PrivBERT and RoBERTa, further indicating that the privacy language domain is distinct and requires its own NLP benchmark.

6.2. Model-pair agreement analysis

PrivBERT, the top performing model, differentiates itself from other models by its in-domain pretraining on the PrivaSeer corpus [27]. Therefore, we can infer that PrivBERT incorporated *knowledge* of privacy policies through its pretraining and became specialized for fine-tuning tasks in the privacy language domain. We investigate this specialization using model-pair agreement analysis to detect examples where PrivBERT had a competitive advantage over other models. Consequently, we detect examples where PrivBERT was disadvantaged due to its in-domain pretraining.

We compare $10 \times 10 = 100$ random seed combinations for all test-set pairs between PrivBERT and other models. Each prediction-pair can be classified into one of four mutually exclusive categories (B, P, O and N) shown below. Categories B and N represent examples that are either not challenging or too challenging for both PrivBERT and the other model respectively. Categories P and O are more interesting for us since they indicate examples where PrivBERT had a competitive advantage and disadvantage over the other model respectively. Therefore, we focus on categories P and O in our analysis. We classify examples over all random seed combinations and take the majority occurrence for each category within its distribution.

Category B: Both PrivBERT and the other model were correct, i.e. (PrivBERT, Other Model)
Category P: PrivBERT was correct and the other model was wrong, i.e. (PrivBERT, ¬ Other Model)
Category O: Other model was correct and PrivBERT was wrong, i.e. (¬ PrivBERT, Other Model)
Category N: Neither PrivBERT nor the other model was correct, i.e. (¬ PrivBERT, ¬ Other Model)

Figure 3 shows a relative distribution of majority categories across model-pairs and PrivacyGLUE tasks. We observe that category P is always greater than category O, which correlates with PrivBERT outperforming all other models. We also observe that category P is often the greatest when compared against BERT, implying that PrivBERT has the most competitive advantage over BERT. Surprisingly, we also observe category O is often the greatest when compared against BERT, implying that BERT has the highest absolute advantage over PrivBERT. This is an insightful observation since we would have expected BERT to have the least competitive advantage given its lowest overall PrivacyGLUE performance.

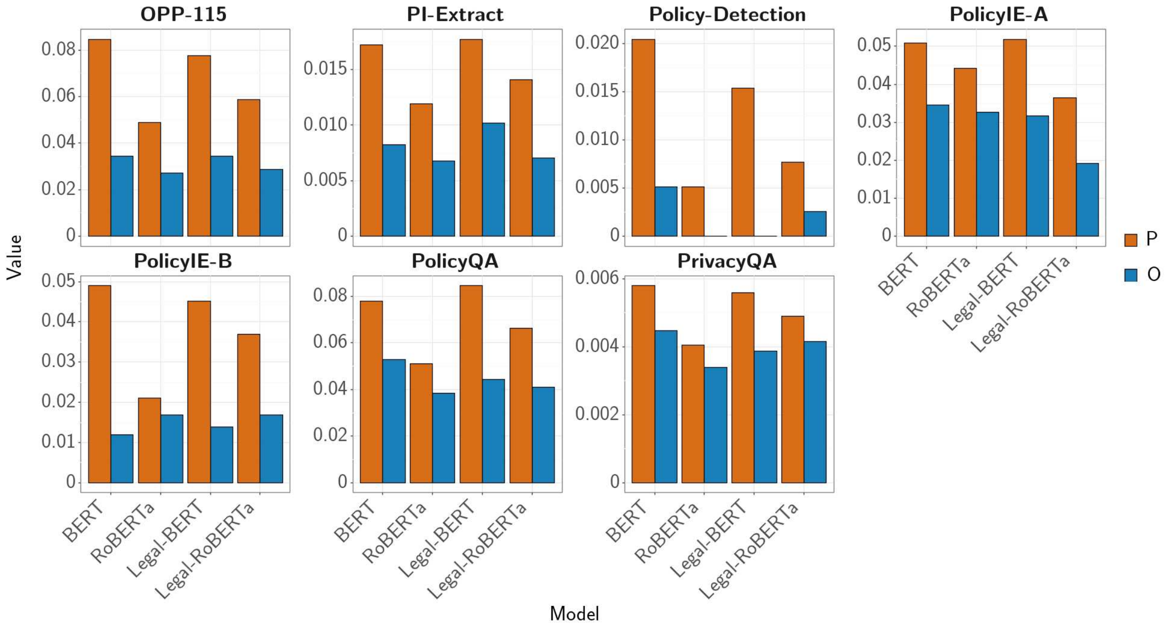

Figure 3. Model-pair agreement analysis of PrivBERT against other models over all PrivacyGLUE tasks; bars represent proportions of examples per model-pair and task which fell into categories P and O; all models on the x-axis are compared against PrivBERT

To investigate PrivBERT's competitive advantage and disadvantage against BERT, we extract several examples from categories P and O in the PrivacyQA task for brevity. Two interesting examples are listed in Table 4 and additional examples can be found in Table A3 in Appendix D. From Table 4, we speculate that PrivBERT specializes in example 1978 because it contains several privacy-specific terms such as "third parties" and "explicit consent". On the other hand, we speculate that BERT specializes in example 33237 since it contains more generic information regarding encryption and SSL, which also happens to be a topic in BERT's Wikipedia pretraining corpus as seen in Figure 1 and Table 2.

Looking at further examples in Table A3, we can also observe that all sampled category P examples have the `Relevant` label while many sampled category O examples have the `Irrelevant` label. On further analysis of the PrivacyQA test-set, we find that 71% of category P examples have the `Relevant` label and 61% of category O samples have the `Irrelevant` label. We can infer that PrivBERT specializes in the minority `Relevant` label while BERT specializes in the majority `Irrelevant` label as the former label could require more privacy knowledge than the latter.

Table 4. Test-set examples from PrivacyQA that fall under categories P and O for PrivBERT vs. BERT

Category P	Category O
ID: 1978 **Question:** Who can see my information? **Answer:** We do not sell or rent your personal information to third parties for their marketing purposes without your explicit consent. **Label:** `Relevant`	**ID:** 33237 **Question:** Could the wordscapes app contain malware? **Answer:** We encrypt the transmission of all information using secure socket layer technology (SSL). **Label:** `Relevant`

7. Conclusions and further work

In this paper, we describe the importance of data privacy in modern digital life and observe the lack of an NLP benchmark in the privacy language domain despite its distinctness. To address this, we propose PrivacyGLUE as the first comprehensive benchmark for measuring general language understanding in the privacy language domain. We release benchmark performances from the BERT, RoBERTa, Legal-BERT, Legal-RoBERTa and PrivBERT transformer language models. Our findings show that PrivBERT outperforms other models by an average of $2-3\%$ over all PrivacyGLUE tasks, shedding light on the importance of in-domain pretraining for privacy policies. We apply model-pair agreement analysis to detect PrivacyGLUE examples where PrivBERT's pretraining provides competitive advantage and disadvantage. By benchmarking holistic model performances, we believe PrivacyGLUE can accelerate NLP research into the privacy language domain and ultimately improve general language understanding of privacy policies for both humans and AI algorithms. Ultimately, this will support practitioners in the practical adoption of NLP models within the privacy domain, for example in assisting consumers with the comprehension of privacy policies in their daily digital lives.

Looking forward, we envision several ways to further enhance our study. Firstly, we intend to apply deep-learning explainability techniques such as Integrated Gradients [33] on examples from Table 4, to explore PrivBERT's and BERT's token-level attention attributions for categories P and O. Additionally, we intend to benchmark large prompt-based transformer language models such as T5 [34] and T0 [35], as they incorporate large amounts of knowledge from the various sequence-to-sequence tasks that they were trained on. Finally, we plan to continue maintaining our PrivacyGLUE GitHub repository and host new model results from the community.

Limitations

To the best of our knowledge, our study has two main limitations. While we provide performances from transformer language models, our study does not provide human expert performances on PrivacyGLUE. This would have been a valuable contribution to judge how competitive language models are against human expertise. However, this limitation can be challenging to address due to the difficulty in finding experts and high costs for their services. Additionally, our study only focuses on English language privacy tasks and omits multilingual scenarios. Multilingual tasks would have been very interesting and relevant to explore, but also involve significant complexity since privacy experts for non-English languages may be harder to find.

Ethics Statement

Original work attribution

All datasets used to compose PrivacyGLUE are publicly available and originate from previous studies. We cite these studies in our paper and include references for them in our GitHub repository. Furthermore, we clearly illustrate how these datasets were used to form the PrivacyGLUE benchmark.

Social impact

PrivacyGLUE could be used to produce fine-tuned transformer language models, which could then be utilized in downstream applications to help users understand privacy policies and/or answer questions regarding them. We believe this could have a positive social impact as it would empower users to better understand lengthy and complex privacy policies. That being said, application developers should perform appropriate risk analyses when using fine-tuned transformer language models. Important points to consider include the varying performance ranges on PrivacyGLUE tasks and known examples of implicit bias, such as gender and racial bias, that transformer language models incorporate through their large-scale pretraining [36].

Software licensing

We release source code for PrivacyGLUE under version 3 of the GNU General Public License (GPL-3.0). We chose GPL-3.0 as it is a strong copyleft license that protects user freedoms such as the freedom to use, modify and distribute software.

Author Contributions: Conceptualization, A.S., A.W, C.B., M.A.R. and L.M.; methodology, A.S., C.B., M.A.R., A.W. and L.M.; software, A.S. and C.B.; validation, A.S., C.B., M.A.R., A.W. and L.M.; formal analysis, A.S. and C.B.; investigation, A.S., C.B., M.A.R., A.W. and L.M.; resources, A.S. and A.W.; data curation, A.S., C.B. and A.W.; writing—original draft preparation, A.S., C.B., M.A.R., A.W. and L.M.; writing—review and editing, A.S., C.B., M.A.R., A.W. and L.M.; visualization, A.W., A.S. and C.B.; supervision, L.M. and A.S.; project administration, L.M.; funding acquisition, L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Repository will be made public post-acceptance. Anonymous repository: https://anonymous.4open.science/r/f4293357886f671347fa69fae3650543

Conflicts of Interest: he authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UN	United Nations
AI	Artificial Intelligence
NLP	Natural Language Processing
GLUE	General Language Understanding Evaluation benchmark
SuperGLUE	Super General Language Understanding Evaluation benchmark
SOTA	State of the Art
LexGLUE	Legal General Language Understanding Evaluation benchmark
EURLEX	Access to European Union law
BERT	Bidirectional Encoder Representations from Transformers
UMAP	Uniform Manifold Approximation and Projection
PrivacyGLUE	Privacy General Language Understanding Evaluation benchmark
OPP-115	Online Privacy Policies, set of 115
PI-Extract	Personal Information Extraction
PolicyIE	Policy Intent Extraction
PolicyQA	Policy Questions and Answers
PrivacyQA	Privacy Questions and Answers
RoBERTa	Robustly Optimized BERT Pretraining Approach
GEM	Natural Language Generation benchmark Metrics
XTREME	Cross-Lingual Transfer Evaluation of Multilingual Encoders
XTREME-R	XTREME Revisited
NER	Named Entity Recognition
SQuAD	Stanford Question Answering Dataset
T5	Text-To-Text Transfer Transformer
T0	T5 for zero-shot

Appendix A. Benchmark configuration

We run PrivacyGLUE benchmark tasks with the following configuration:

- We train all models for 20 epochs with a batch size of 16. We utilize a linear learning rate scheduler with a warmup ratio of 0.1 and peak learning rate of 3×10^{-5}. We utilize AdamW [37] as our optimizer. Finally, we monitor respective metrics on the validation datasets and utilize early stopping if the validation metric does not improve for 5 epochs.
- We use `Python v3.8.13, CUDA v11.7, PyTorch v1.12.1` [38] and `Transformers v4.19.4` [39] as our core software dependencies.
- We use the following HuggingFace model tags: `bert-base-uncased`, `roberta-base`, `nlpaueb/legal-bert-base-uncased`, `saibo/legal-roberta-base`, `mukund/privbert` for BERT, RoBERTa, Legal-BERT, Legal-RoBERTa and PrivBERT respectively.
- We use 10 random seeds for each benchmark run, i.e. $\{0,1,2,3,4,5,6,7,8,9\}$. This provides a distribution of results that can be used for statistical significance testing.
- We run the PrivacyGLUE benchmark on a Lambda workstation with 4 × NVIDIA RTX A4000 (16 GB VRAM) GPUs for ∼180 hours.
- We use `Weights and Biases v0.13.3` [40] to monitor model metrics during training and for intermediate report generation.

Appendix B. Detailed label information

Table A1. Breakdown of labels for each PrivacyGLUE task; PolicyQA is omitted from this table since it is a reading comprehension task and does not have explicit labels like other tasks

Task	Labels
OPP-115	`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`
PI-Extract	**Subtask-I:** `{B,I}-COLLECT, O` **Subtask-II:** `{B,I}-NOT_COLLECT, O` **Subtask-III:** `{B,I}-NOT_SHARE, O` **Subtask-IV:** `{B,I}-SHARE, O`
Policy-Detection	`Not Policy, Policy`
PolicyIE-A	`Other, data-collection-usage, data-security-protection, data-sharing-disclosure, data-storage-retention-deletion`
PolicyIE-B	**Subtask-I:** `{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, storage-place, {B,I}-retention-period, {B,I}-protect-against, {B,I}-action, O` **Subtask-II:** `{B,I}-purpose-argument, {B,I}-polarity, {B,I}-method, {B,I}-condition-argument, O`
PrivacyQA	`Irrelevant, Relevant`

Appendix C. PrivacyGLUE task examples

Table A2. Representative examples of each PrivacyGLUE benchmark task

Task	Input	Target
OPP-115	Revision Date: March 24th 2015	Introductory/Generic, Policy Change
PI-Extract	We may collect and share your IP address but not your email address with our business partners .	**Subtask-I:** O O O O O B-COLLECT I-COLLECT I-COLLECT O O O O O O O O O O **Subtask-II:** 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-II:** 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:** O O O O O B-SHARE I-SHARE I-SHARE O O O O O O O O O O
Policy-Detection	Log in through another service: * Facebook * Google	Not Policy
PolicyIE-A	To backup and restore your Pocket AC camera log	data-collection-usage
PolicyIE-B	Access to your personal information is restricted .	**Subtask-I:** O O B-data-provider B-data-protected I-data-protected O B-action O **Subtask-II:** B-method O O O O O O O
PolicyQA	**Question:** How do they secure my data? **Context:** Users can visit our site anonymously	**Answer:** Users can visit our site anonymously
PrivacyQA	**Question:** What information will you collect about my usage? **Answer:** Location information	Relevant

Appendix D. Additional PrivacyQA examples from categories P and O

Table A3. Additional test-set examples from PrivacyQA that fall under categories P and O for PrivBERT vs. BERT; note that these examples are not paired and can therefore be compared in any order between categories

Category P	Category O
ID: 9227 **Question:** Will the app use my data for marketing purposes? **Answer:** We will never share with or sell the information gained through the use of Apple HealthKit, such as age, weight and heart rate data, to advertisers or other agencies without your authorization. **Label:** Relevant	**ID:** 8749 **Question:** Will my fitness coach share my information with others? **Answer:** Develop new services. **Label:** Irrelevant
ID: 10858 **Question:** What information will this app have access to of mine? **Answer:** Information you make available to us when you open a Keep account, as set out above; **Label:** Relevant	**ID:** 47271 **Question:** Who will have access to my medical information? **Answer:** 23andMe may share summary statistics, which do not identify any particular individual or contain individual-level information, with our qualified research collaborators. **Label:** Irrelevant

Table A3. *Cont.*

Category P	**Category O**
ID: 18704 **Question:** Does it share my personal information with others? **Answer:** We may also disclose Non-Identifiable Information: **Label:** `Relevant`	**ID:** 54904 **Question:** What data do you keep and for how long? **Answer:** We may keep activity data on a non-identifiable basis to improve our services. **Label:** `Irrelevant`
ID: 45935 **Question:** Will my test results be shared with any third party entities? **Answer:** 23andMe may share summary statistics, which do not identify any particular individual or contain individual-level information, with our qualified research collaborators. **Label:** `Relevant`	**ID:** 57239 **Question:** Do you sell any of our data? **Answer:** (c) Advertising partners: to enable the limited advertisements on our service, we may share a unique advertising identifier that is not attributable to you, with our third party advertising partners, and advertising service providers, along with certain technical data about you (your language preference, country, city, and device data), based on our legitimate interest. **Label:** `Relevant`
ID: 50467 **Question:** Can I delete my personally identifying information? **Answer:** (Account Deletion), we allow our customers to delete their accounts at any time. **Label:** `Relevant`	**ID:** 59334 **Question:** Does the app protect my account details from being accessed by other people? **Answer:** Note that chats with bots and Public Accounts, and communities are not end-to-end encrypted, but we do encrypt such messages when sent to the Viber servers and when sent from the Viber servers to the third party (the Public Account owner and/or additional third party tool (eg CRM solution) integrated by such owner). **Label:** `Irrelevant`

Appendix E. PrivacyGLUE benchmark results

Table A4. Test-set results of the PrivacyGLUE benchmark; † m-F$_1$ refers to macro-average F$_1$, ¯-F$_1$ refers to the micro-average F$_1$, s refers to sample-average F$_1$, EM refers to the exact match accuracy, metrics are reported as percentages with the following format: mean$_{\pm\text{standard deviation}}$

Task	**Metric**[†]	**BERT**	**RoBERTa**	**Legal-BERT**	**Legal-RoBERTa**	**PrivBERT**
OPP-115	m-F$_1$	78.4$_{\pm 0.6}$	79.5$_{\pm 1.1}$	79.6$_{\pm 1.0}$	79.1$_{\pm 0.7}$	**82.1**$_{\pm 0.5}$
	¯-F$_1$	84.0$_{\pm 0.5}$	85.4$_{\pm 0.5}$	84.3$_{\pm 0.7}$	84.7$_{\pm 0.3}$	**87.2**$_{\pm 0.4}$
PI-Extract	m-F$_1$	60.0$_{\pm 2.7}$	62.4$_{\pm 4.4}$	59.5$_{\pm 3.0}$	60.5$_{\pm 3.9}$	**66.4**$_{\pm 3.4}$
	¯-F$_1$	60.0$_{\pm 2.7}$	62.4$_{\pm 4.4}$	59.5$_{\pm 3.0}$	60.5$_{\pm 3.9}$	**66.4**$_{\pm 3.4}$
Policy-Detection	m-F$_1$	85.3$_{\pm 1.8}$	86.9$_{\pm 1.3}$	86.6$_{\pm 1.0}$	86.4$_{\pm 2.0}$	**87.3**$_{\pm 1.1}$
	¯-F$_1$	92.1$_{\pm 1.2}$	92.7$_{\pm 0.8}$	92.7$_{\pm 0.5}$	92.4$_{\pm 1.3}$	**92.9**$_{\pm 0.8}$
PolicyIE-A	m-F$_1$	72.9$_{\pm 1.7}$	73.2$_{\pm 1.6}$	73.2$_{\pm 1.5}$	73.5$_{\pm 1.5}$	**75.3**$_{\pm 2.2}$
	¯-F$_1$	84.7$_{\pm 1.0}$	84.8$_{\pm 0.6}$	84.7$_{\pm 0.5}$	84.8$_{\pm 0.3}$	**86.2**$_{\pm 1.0}$
PolicyIE-B	m-F$_1$	50.3$_{\pm 0.7}$	52.8$_{\pm 0.6}$	51.5$_{\pm 0.7}$	53.5$_{\pm 0.5}$	**55.4**$_{\pm 0.7}$
	¯-F$_1$	50.3$_{\pm 0.5}$	54.5$_{\pm 0.7}$	52.2$_{\pm 1.0}$	53.6$_{\pm 0.9}$	**55.7**$_{\pm 1.3}$
PolicyQA	s-F$_1$	55.7$_{\pm 0.5}$	57.4$_{\pm 0.4}$	55.3$_{\pm 0.7}$	56.3$_{\pm 0.6}$	**59.3**$_{\pm 0.5}$
	EM	28.0$_{\pm 0.9}$	30.0$_{\pm 0.5}$	27.5$_{\pm 0.6}$	28.6$_{\pm 0.9}$	**31.4**$_{\pm 0.6}$
PrivacyQA	m-F$_1$	53.6$_{\pm 0.8}$	54.4$_{\pm 0.3}$	53.6$_{\pm 0.8}$	54.4$_{\pm 0.5}$	**55.3**$_{\pm 0.6}$
	¯-F$_1$	90.0$_{\pm 0.1}$	90.2$_{\pm 0.0}$	90.0$_{\pm 0.1}$	90.2$_{\pm 0.1}$	**90.2**$_{\pm 0.1}$

doi:10.20944/preprints202303.0046.v1

References

1. Gstrein, O.J.; Beaulieu, A. How to protect privacy in a datafied society? A presentation of multiple legal and conceptual approaches. *Philosophy & Technology* **2022**, *35*, 1–38.
2. Obar, J.A.; Oeldorf-Hirsch, A. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* **2020**, *23*, 128–147.
3. McDonald, A.M.; Cranor, L.F. The cost of reading privacy policies. *Isjlp* **2008**, *4*, 543.
4. Wilson, S.; Schaub, F.; Dara, A.A.; Liu, F.; Cherivirala, S.; Giovanni Leon, P.; Schaarup Andersen, M.; Zimmeck, S.; Sathyendra, K.M.; Russell, N.C.; Norton, T.B.; Hovy, E.; Reidenberg, J.; Sadeh, N. The Creation and Analysis of a Website Privacy Policy Corpus. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics: Berlin, Germany, 2016; pp. 1330–1340. doi:10.18653/v1/P16-1126.
5. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 353–355. doi:10.18653/v1/W18-5446.
6. Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. Advances in Neural Information Processing Systems; Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; Garnett, R., Eds. Curran Associates, Inc., 2019, Vol. 32.
7. Chalkidis, I.; Jana, A.; Hartung, D.; Bommarito, M.; Androutsopoulos, I.; Katz, D.; Aletras, N. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics: Dublin, Ireland, 2022; pp. 4310–4330. doi:10.18653/v1/2022.acl-long.297.
8. Wikimedia Foundation. Wikimedia Downloads, 2022.
9. Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Androutsopoulos, I. Large-Scale Multi-Label Text Classification on EU Legislation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Florence, Italy, 2019; pp. 6314–6322. doi:10.18653/v1/P19-1636.
10. Mazzola, L.; Waldis, A.; Shankar, A.; Argyris, D.; Denzler, A.; Van Roey, M. Privacy and Customer's Education: NLP for Information Resources Suggestions and Expert Finder Systems. HCI for Cybersecurity, Privacy and Trust; Moallem, A., Ed.; Springer International Publishing: Cham, 2022; pp. 62–77.
11. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints* **2018**, [arXiv:stat.ML/1802.03426].
12. Gehrmann, S.; Adewumi, T.; Aggarwal, K.; Ammanamanchi, P.S.; Aremu, A.; Bosselut, A.; Chandu, K.R.; Clinciu, M.A.; Das, D.; Dhole, K.; Du, W.; Durmus, E.; Dušek, O.; Emezue, C.C.; Gangal, V.; Garbacea, C.; Hashimoto, T.; Hou, Y.; Jernite, Y.; Jhamtani, H.; Ji, Y.; Jolly, S.; Kale, M.; Kumar, D.; Ladhak, F.; Madaan, A.; Maddela, M.; Mahajan, K.; Mahamood, S.; Majumder, B.P.; Martins, P.H.; McMillan-Major, A.; Mille, S.; van Miltenburg, E.; Nadeem, M.; Narayan, S.; Nikolaev, V.; Niyongabo Rubungo, A.; Osei, S.; Parikh, A.; Perez-Beltrachini, L.; Rao, N.R.; Raunak, V.; Rodriguez, J.D.; Santhanam, S.; Sedoc, J.; Sellam, T.; Shaikh, S.; Shimorina, A.; Sobrevilla Cabezudo, M.A.; Strobelt, H.; Subramani, N.; Xu, W.; Yang, D.; Yerukola, A.; Zhou, J. The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021); Association for Computational Linguistics: Online, 2021; pp. 96–120. doi:10.18653/v1/2021.gem-1.10.
13. Hu, J.; Ruder, S.; Siddhant, A.; Neubig, G.; Firat, O.; Johnson, M. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. Proceedings of the 37th International Conference on Machine Learning; III, H.D.; Singh, A., Eds. PMLR, 2020, Vol. 119, *Proceedings of Machine Learning Research*, pp. 4411–4421.
14. Ruder, S.; Constant, N.; Botha, J.; Siddhant, A.; Firat, O.; Fu, J.; Liu, P.; Hu, J.; Garrette, D.; Neubig, G.; Johnson, M. XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Online and Punta Cana, Dominican Republic, 2021; pp. 10215–10245. doi:10.18653/v1/2021.emnlp-main.802.

15. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 4171–4186. doi:10.18653/v1/N19-1423.
16. Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; Androutsopoulos, I. LEGAL-BERT: The Muppets straight out of Law School. Findings of the Association for Computational Linguistics: EMNLP 2020; Association for Computational Linguistics: Online, 2020; pp. 2898–2904. doi:10.18653/v1/2020.findings-emnlp.261.
17. Zimmeck, S.; Story, P.; Smullen, D.; Ravichander, A.; Wang, Z.; Reidenberg, J.R.; Russell, N.C.; Sadeh, N. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.* **2019**, *2019*, 66.
18. Bui, D.; Shin, K.G.; Choi, J.M.; Shin, J. Automated Extraction and Presentation of Data Practices in Privacy Policies. *Proceedings on Privacy Enhancing Technologies* **2021**, *2021*, 88–110. doi:doi:10.2478/popets-2021-0019.
19. Ahmad, W.; Chi, J.; Le, T.; Norton, T.; Tian, Y.; Chang, K.W. Intent Classification and Slot Filling for Privacy Policies. 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); Association for Computational Linguistics: Online, 2021; pp. 4402–4417. doi:10.18653/v1/2021.acl-long.340.
20. Ravichander, A.; Black, A.W.; Wilson, S.; Norton, T.; Sadeh, N. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Association for Computational Linguistics: Hong Kong, China, 2019; pp. 4949–4959. doi:10.18653/v1/D19-1500.
21. Ahmad, W.; Chi, J.; Tian, Y.; Chang, K.W. PolicyQA: A Reading Comprehension Dataset for Privacy Policies. Findings of the Association for Computational Linguistics: EMNLP 2020; Association for Computational Linguistics: Online, 2020; pp. 743–749. doi:10.18653/v1/2020.findings-emnlp.66.
22. Amos, R.; Acar, G.; Lucherini, E.; Kshirsagar, M.; Narayanan, A.; Mayer, J. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. Proceedings of The Web Conference 2021. Association for Computing Machinery, 2021, WWW '21, p. 22. doi:10.1145/3442381.3450048.
23. Mousavi Nejad, N.; Jabat, P.; Nedelchev, R.; Scerri, S.; Graux, D. Establishing a strong baseline for privacy policy classification. IFIP International Conference on ICT Systems Security and Privacy Protection. Springer, 2020, pp. 370–383.
24. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Austin, Texas, 2016; pp. 2383–2392. doi:10.18653/v1/D16-1264.
25. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**.
26. Geng, S.; Lebret, R.; Aberer, K. Legal Transformer Models May Not Always Help. *CoRR* **2021**, *abs/2109.06862*, [2109.06862].
27. Srinath, M.; Wilson, S.; Giles, C.L. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. 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); Association for Computational Linguistics: Online, 2021; pp. 6829–6839. doi:10.18653/v1/2021.acl-long.532.
28. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; others. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* **2016**.
29. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics: Berlin, Germany, 2016; pp. 1715–1725. doi:10.18653/v1/P16-1162.
30. Kudo, T.; Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 66–71. doi:10.18653/v1/D18-2012.

31. Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* **1947**, pp. 50–60.
32. Dror, R.; Shlomov, S.; Reichart, R. Deep Dominance - How to Properly Compare Deep Neural Models. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Florence, Italy, 2019; pp. 2773–2785. doi:10.18653/v1/P19-1266.
33. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. *CoRR* **2017**, *abs/1703.01365*, [1703.01365].
34. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* **2020**, *21*, 1–67.
35. Sanh, V.; Webson, A.; Raffel, C.; Bach, S.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Raja, A.; Dey, M.; Bari, M.S.; Xu, C.; Thakker, U.; Sharma, S.S.; Szczechla, E.; Kim, T.; Chhablani, G.; Nayak, N.; Datta, D.; Chang, J.; Jiang, M.T.J.; Wang, H.; Manica, M.; Shen, S.; Yong, Z.X.; Pandey, H.; Bawden, R.; Wang, T.; Neeraj, T.; Rozen, J.; Sharma, A.; Santilli, A.; Fevry, T.; Fries, J.A.; Teehan, R.; Scao, T.L.; Biderman, S.; Gao, L.; Wolf, T.; Rush, A.M. Multitask Prompted Training Enables Zero-Shot Task Generalization. International Conference on Learning Representations, 2022.
36. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery: New York, NY, USA, 2021; FAccT '21, p. 610–623. doi:10.1145/3442188.3445922.
37. Loshchilov, I.; Hutter, F. Fixing Weight Decay Regularization in Adam. *CoRR* **2017**, *abs/1711.05101*, [1711.05101].
38. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems; Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; Garnett, R., Eds. Curran Associates, Inc., 2019, Vol. 32.
39. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T.L.; Gugger, S.; Drame, M.; Lhoest, Q.; Rush, A.M. Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Association for Computational Linguistics: Online, 2020; pp. 38–45.
40. Biewald, L. Experiment Tracking with Weights and Biases, 2020. Software available from wandb.com.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.