A Systematic Evaluation of Model Architectures for Privacy Policy Question Answering

Abhas Wanchu
University of California, Berkeley
Masters of Information and Data Science
Course: DATASCI266
abhas@berkeley.edu
https://github.com/abhaswanchu1/mids-266-final-project.git

Abstract

This paper presents a comprehensive methodology for fine-tuning and evaluating various language models on the task of question answering (QA) within the legal and policy domain. We specifically investigate the performance of a fine-tuned OpenAI GPT-4.1-Nano, a Legal-BERT model, and a RoBERTa-LARGE model. Additionally, we explore the feasibility of implementing a Retrieval-Augmented Generation (RAG) system through limited preliminary tests. The research explores the impact of few-shot learning, system message variation, and hyperparameter tuning on model performance, measured by Precision, Recall, F1, and Exact Match (EM) scores. Our findings provide valuable insights into the capabilities and limitations of different models and prompting strategies for legal QA, highlighting the potential for building robust and trustworthy domain-specific systems.

1 Introduction

The legal and policy domain poses significant challenges for question answering systems due to the complexity and specialized terminology of legal texts (3; 5). While Large Language Models (LLMs) have shown general QA capabilities, their application in this domain necessitates careful fine-tuning for accuracy and reliability. This research details experiments addressing these challenges by fine-tuning pre-trained models and developing a RAG system for policy documents. We focus on the PolicyQA (1) and PrivacyQA datasets from the PLUE (7) Benchmarks, employing a systematic approach to model evaluation and hyperparameter optimization. This paper outlines the methodologies, preliminary findings, and future research directions for legal AI.

1.1 Motivation: The Critical Need for AI-Driven Privacy Policy Understanding

Existing literature on data privacy is predicated on the assumption that consumers are provided with a clear notice and subsequently provide consent for the collection and use of their personal data. However, despite 97% of Americans reporting that they are asked to approve privacy policies, only one-in-five adults (22%) consistently read these documents, with 36% admitting they never do. Furthermore, a significant majority (57%) express a lack of confidence that companies adhere to their stated privacy policies (6).

This demonstrates how users neglect lengthy and complex privacy policies, leading to a transparency gap that undermines informed consent and control over personal information. This demands innovative AI solutions, particularly NLP-driven QA systems, to facilitate quick and accurate understanding of privacy practices. Such tools can transform opaque legal documents into accessible insights, aligning with responsible AI development principles of fairness, transparency, and accountability. Developing AI systems capable of accurate and ethical interpretation of privacy policies is crucial for enhancing digital literacy and user autonomy.

1.2 Literature Review: Privacy Policy Understanding and Question Answering

The increasing complexity of online privacy policies necessitates advanced NLP solutions, particularly QA systems, to assist users in understanding their data rights. Early efforts led to foundational datasets like PrivacyQA, which highlighted challenges such as class imbalance, followed by PolicyQA (1), which focused on extracting precise answer spans. The multi-task PLUE (7) benchmark further broadened the scope of evaluation for privacy policy language understanding. These developments underscore a continuous drive for more realistic and challenging environments to assess

NLP models in this domain.

A consistent finding is the critical role of domain-specific pre-training for high performance in legal and privacy NLP tasks (3; 5). Studies confirm that models adapted to legal or privacy-related corpora significantly outperform general-purpose models, indicating that the unique linguistic characteristics of legal texts require specialized adaptation. For instance, models continually pre-trained on privacy policy corpora demonstrate noticeable performance improvements across PLUE tasks, with PP-RoBERTaLARGE and RoBERTaLARGE showing the highest F1 scores in PolicyQA, and PP-RoBERTaBASE and PP-ElectraBASE excelling in PrivacyQA (7). These findings establish that while large, general-purpose language models provide a strong foundation, their full potential in specialized domains like privacy policy understanding is unlocked through targeted pre-training or fine-tuning on relevant legal and policy texts.

Despite these advancements, significant research gaps persist, including the need for models that can handle dynamic privacy policies, provide explainable and trustworthy answers, facilitate cross-lingual transfer, and perform more fine-grained analysis beyond simple relevance classification. Addressing these areas is crucial for developing truly robust and practical privacy policy QA systems that empower users with actionable insights.

Our research centers on two datasets from the PLUE Benchmarks: PolicyQA and PrivacyQA.

PolicyQA: This dataset is formulated as a span detection task, requiring models to identify and extract the precise text span within a policy document that contains the answer to a specific question.

- **Exploratory Data Analysis (EDA) of the PolicyQA Training Data:** The training dataset comprises 75 documents broken down into 2,189 paragraphs, yielding 17,056 question-answer pairs. Each tokenized example corresponds to a single question-paragraph pair. A total of 26,861 answer lengths were extracted, indicating multiple answer spans for some questions. Answer lengths vary significantly (1 to 289 words, mean 11.44, median 5, right-skewed). Context lengths range from 24 to 2,279 characters (mean 471.6, median 405, right-skewed). Question lengths are more consistent (5 to 19 words, mean 9.58, median 9, roughly bell-shaped). Answer start positions range from 0 to 2,201 (mean 225, median 147, skewed towards the beginning of paragraphs). These findings necessitate robust tokenization strategies (e.g., truncation with stride) to prevent information loss. While answers tend to appear earlier, the wide range of starting positions means models must process the entire context. The consistent question lengths suggest less challenge in question processing. The presence of multiple answer spans is a crucial characteristic for models to handle during training and inference. Both validation and test datasets contained 20 documents each, with a varying number of paragraphs and question-answer pairs

PrivacyQA: This dataset is structured as a binary sequence classification task, where the model classifies a text sequence as either "relevant" or "irrelevant" to a given question, aiding in noise reduction and focusing on pertinent information.

- **PrivacyQA Training Dataset Challenges:** This dataset presents two primary challenges: severe class imbalance and significant variation in text lengths. The most critical finding is the overwhelming 96.16% (160,280 examples) of "Irrelevant" labels versus only 3.84% (6,400 examples) of "Relevant" labels. This imbalance poses a major hurdle, potentially biasing models toward predicting the majority class. Additionally, query text lengths range from 10 to 126 characters, while segment text entries vary widely from 1 to 1313 characters, necessitating careful tokenization and input formatting. While common keywords confirm the dataset's privacy focus, the analysis shows no strong correlation between text length and segment relevance, suggesting successful classification requires complex linguistic features. 10% of the training set was held out for validation, and the test set contained 3,137 examples.

2 Methodology

The research methodology comprises three primary components: fine-tuning and evaluation of Legal-BERT-Small, RoBERTa-Large, and OpenAI GPT-4.1-nano models.

2.1 PolicyQA

2.1.1 Legal-BERT Fine-tuning and RAG Implementation

For the Legal-BERT model, the PolicyQA dataset was tokenized and formatted for answer span identification. An out-of-the-box BERT model was used to develop a baseline. Hyperparameter tuning experiments explored learning rates (1e-5, 3e-5, 5e-5), training epochs (2, 5, 10), maximum sequence length, stride, batch size, and weight decay. Each configuration was run three times, with average F1 and EM scores identifying the best setup. A final model was trained with optimal parameters and evaluated on development and test sets.

A preliminary experiment integrated the fine-tuned Legal-BERT model into a RAG system. This involved generating embeddings for policy contexts and building a FAISS vector index. The RAG system retrieved relevant contexts for a given question, passing them to OpenAI's GPT-4.1-Nano for answer generation. The end-to-end RAG system was evaluated on the development set using SQuAD metrics.

2.1.2 LLM Evaluation and One/Few-Shot Learning

This research also evaluated LLMs on the PolicyQA task, specifically examining few-shot learning and system message variations. The primary model evaluated was a fine-tuned OpenAI GPT-4.1-Nano. This model was fine-tuned for the PLUE benchmark task. Experiments used few-shot learning by providing a small number of question-answer examples (1, 2, 3, 5, or 10 shots). System message variation was explored using instructions such as "Strict Extractive" or "Concise Answer." The evaluation involved prompting the model with few-shot examples and test queries, comparing predicted answers to ground truth using F1 and EM scores.

2.1.3 RoBERTa-Large Fine-Tuning

This section details the methodology for developing a Question Answering (QA) system for policy documents using a fine-tuned RoBERTa-Large model within a Retrieval-Augmented Generation (RAG) framework.

Data Preparation: The custom policy document dataset was loaded and preprocessed using the appropriate tokenizer; this included handling long contexts via truncation and overlapping windows (stride) and mapping ground truth answer spans to token indices.

Model Fine-tuning: The RoBERTa model underwent supervised fine-tuning using the Hugging Face Trainer API. Systematic hyperparameter tuning explored learning rates (e.g., 1e-5 to 1e-4), training epochs (e.g., 1-10), batch sizes (e.g., 8-64), and a weight decay of 0.01. Multiple runs (typically three) were executed for each configuration, with average Exact Match (EM) and F1 Score on the development set guiding the selection of optimal hyperparameters: Learning Rate 3e-5, 5 Epochs, Batch Size 16, and Weight Decay 0.01.

Evaluation Metrics: Exact Match (EM) and F1-score were consistently used across all architectures; for fine-tuned models, we fine-tune them three times to report average performances. The complete results of hyperparameter tuning and experiments is available in our code repository.

2.2 PrivacyQA

Addressing Class Imbalance with Downsampling: Due to the severe class imbalance in PrivacyQA, the majority "Irrelevant" class in the training dataset was randomly undersampled to match the number of "Relevant" examples, creating a balanced dataset to mitigate bias.

2.2.1 RoBERTa-large for Binary Classification

A pre-trained 'RoBERTa-large' model was then fine-tuned on this balanced dataset. The model was evaluated on a validation set and tuned using random search.

2.2.2 Legal-BERT-Small for Binary Classification

A pre-trained Legal-BERT-Small model was fine tuned on the balanced dataset. The model was evaluated on a validation data and tuned via random search.

Hyperparameter Tuning: A random search strategy was employed on the development set to optimize hyperparameters. The search space included learning rates (5e-5, 3e-5, 2e-5), per-device train batch size, epochs (2, 3, 5, 10), and weight decay (0.0, 0.01). Ten random trials were conducted. For each trial, the model was re-initialized, configured with trial-specific TrainingArguments, and trained. Performance was evaluated on the development set, and F1-score was the primary metric for selecting optimal hyperparameters.

2.2.3 LLM Evaluation and One/Few-Shot Learning

This section details few-shot learning experiments using GPT-4.1-Nano for binary sentence classification on the PrivacyQA test dataset, leveraging its in-context learning.

Few-Shot Example Selection: Representative examples (three "Relevant" and three "Irrelevant") were randomly selected from the PrivacyQA training dataset to serve as in-context demonstrations.

Prompt Construction for Few-Shot Learning: Key elements included:

1. **System Instruction:** "Act as a legal expert and classify the following sentences as 'Relevant' or 'Irrelevant' to privacy policies."

2. **In-Context Examples:** Relevant and irrelevant examples were included based on the desired number of shots (1, 2, or 3), formatted as "Sentence: [text]\nLabel: [label]". An odd number of shots included one more relevant example.

3. **Class Imbalance Note:** "Note: The dataset has a significant class imbalance, with many more 'Irrelevant' examples than 'Relevant' ones."

4. **Test Sentence:** The test sentence was appended, followed by "Label:" for prediction.

Prediction Process: Experiments were conducted for 1-shot, 2-shot, and 3-shot configurations. For each, the system iterated through the PrivacyQA test dataset, generated a prompt, sent it to the model via the OpenAI API, and extracted the predicted label. Predicted labels for each shot configuration were stored. The best-shot model performance on the validation set was used for final evaluation.

Evaluation Metrics: Precision, Recall, and F1-score were consistently used across all architectures, and we fine-tune all the models three times to report average performances. The complete results of hyperparameter tuning and experiments is available in our code repository.

3 Results and Discussion

3.1 PrivacyQA

This section presents the empirical results for models evaluated on the PrivacyQA task, a binary sentence classification task. We compare the performance of our models using Precision, Recall, and F1-score.

3.1.1 Results

Table 1 summarizes model performance on the PrivacyQA binary sentence classification task, presenting both overall and class-wise metrics, with emphasis on the "Relevant" class.

article

3.1.2 Discussion

The evaluation of models on the PrivacyQA task, particularly focusing on the "Relevant" class due to the dataset's class imbalance, reveals significant performance differences.

Both the TF-IDF + Logistic Regression Baseline and Simple Embedding (Bag-of-Words + Linear SVC) Baseline models achieve decent weighted average F1-scores (0.78 and 0.75, respectively). However, their "Relevant" class F1-scores (0.21 and 0.20) are poor, indicating a struggle to identify actual relevant sentences despite high performance on the majority "Irrelevant" class.

The Legal-BERT-Small outputs the highest F1 score for this task. The model struggles with the class imbalance present in the test set, and effectively classifies everything as "Irrelevant".

The RoBERTa-Large model exemplifies the impact of class imbalance. While its weighted average F1-score (0.8527) and recall (0.90) appear strong, its class-wise metrics show 0.0 Precision, Recall, and F1-score for the "Relevant" class, alongside perfect recall (1.0) for the "Irrelevant" class. This means it effectively classifies all instances as "Irrelevant," rendering it useless for identifying relevant sentences and highlighting the danger of relying solely on aggregated metrics in imbalanced datasets.

The GPT-4.1-Nano model (2-shot) exhibits a different behavior. Its overall metrics are relatively low (F1-score 0.3715, recall 0.2671). However, its "Relevant" class performance shows perfect recall (1.0000) at the cost of extremely low precision (0.0515), leading to a very low "Relevant" F1-score (0.0979). This indicates gpt-4.1-nano over-predicts "Relevant" to capture all true positive examples, resulting in many false positives.

3.2 PolicyQA

This section presents the empirical results from evaluating various models on the PolicyQA task,

Model	Metric	Relevant Class	Irrelevant Class	Weighted Average
Human*	F1-Score			68.9
	Recall			69.0
	Precision			68.8
Baseline:TF-IDF+Logistic Regression	Precision	0.15	0.92	0.84
	Recall	0.35	0.79	0.74
	F1-score	0.21	0.85	0.78
Baseline:Simple Embedding (BOW+ LinearSVC)	Precision	0.14	0.91	0.84
	Recall	0.39	0.72	0.69
	F1-score	0.20	0.81	0.75
Legal-BERT-Small	Precision	0.0	0.92	0.84
	Recall	0.0	1.00	0.92
	F1-score	0.0	0.95	0.88
RoBERTa-Large	Precision	0.0	0.90	0.81
	Recall	0.0	1.0	0.90
	F1-score	0.0	0.94	0.85
GPT-4.1-Nano (2-Shot)	Precision	0.05	1.00	0.96
	Recall	1.00	0.23	0.26
	F1-score	0.09	0.38	0.37

Table 1: Model Performance on PrivacyQA Binary Sentence Classification Task

an extractive Question Answering (QA) span detection task. We compare a BERT baseline, RoBERTa-Large, Legal-BERT-small, and GPT-4.1-nano using F1-score and Exact Match (EM). The task requires models to extract the precise answer span from a policy document in response to a query. EM measures exact matches, while F1-score considers partial overlaps, balancing precision and recall for span detection.

3.2.1 Results

Table 2 summarizes model performance on the PolicyQA span detection QA task.

Model	F1-score (Overall)	Exact Match (EM)
SOTA: ElectraLarge	60.7	33.2
Baseline: BERT*	53.86	27.02
Legal-BERT-small	54.90	27.26
RoBERTa-Large	59.88	32.57
GPT-4.1-Nano(1-Shot)	45.48	10.98

Table 2: Model Performance on PolicyQA Span Detection QA Task

3.2.2 Discussion

The results for the PolicyQA span detection QA task provide insights into how different language models perform in pinpointing answers within legal texts.

The Baseline - BERT (not fine tuned) model (F1: 53.86, EM: 27.02) demonstrates some inherent capability but struggles with exact answer boundaries. The Legal-BERT-Small model shows only a marginal improvement (F1: 54.90, EM: 27.26) over the baseline, suggesting that while domain-specific pre-training offers a slight advantage, it doesn't improve greatly on precise span extraction.

The RoBERTa-Large model achieves the strongest performance (F1: 59.88, EM: 32.57). Its larger architecture and fine-tuning on relevant data likely contribute to its superior ability to identify precise answer spans, and achieves results close to SOTA.

Error analysis for PolicyQA reveals two primary

categories of incorrect predictions: issues with answer span length and difficulty with complex sentence structures. Nearly half of all incorrect predictions (48.23%) occur when the model's predicted answer span is shorter than the reference, with an average length difference of 6.98 tokens. A notable portion (22.87%) of these short predictions are more than 5 tokens shorter. Manual inspection indicates that errors related to complex sentence structures involve misinterpreting negations, struggling with multi-sentence reasoning, failing to identify correct spans in convoluted sentences, and misinterpreting the scope of conditions or exceptions. Further linguistic analysis and manual categorization would be beneficial for a more precise quantification of these nuanced errors.

4 Discussion and Future Work

Our findings underscore the dual nature of AI application in privacy policy understanding: immense potential alongside significant challenges. The severe class imbalance in the PrivacyQA dataset emerged as a critical impediment, leading to models that either entirely predicted the majority class or demonstrated low precision, highlighting the necessity for robust data handling and model training strategies tailored to imbalanced distributions. Additionally, the scarcity of labeled datasets poses a challenge. Future work can explore the incorporation of data augmentation techniques introduced by Parvez et. al (4). For the PolicyQA span detection task, RoBERTa-Large exhibited superior performance, reinforcing the efficacy of larger, finely-tuned transformer architectures in extractive QA. Preliminary RAG investigations also indicate a promising direction for integrating retrieval capabilities with generative models to produce more comprehensive answers.

Future research will primarily address the identified limitations. We also plan to investigate dynamic threshold optimization to achieve a better balance between precision and recall. For LLMs like GPT-4.1-Nano, more sophisticated fine-tuning strategies will be explored beyond few-shot prompting to enhance accuracy and reliability. In the context of PolicyQA and RAG, we will develop more intelligent legal text chunking methods, possibly integrating legal ontologies, and explore hybrid retrieval models combining vector similarity with keyword-based approaches. Furthermore, training domain-specific embeddings will be crucial for capturing the nuances of legal language. Finally, ensemble methods will be investigated to combine the strengths of different models across tasks, aiming for increased robustness and overall performance.

The deployment of AI tools for privacy policy QA and summarization presents both opportunities and profound ethical and legal considerations. While such tools can empower users by making complex policies accessible, thus fostering informed consent and digital literacy, the risks are substantial. Inaccurate or oversimplified summaries could lead to misinterpretations, exposing users to unforeseen legal liabilities or privacy breaches. Questions of accountability arise when AI-generated information proves faulty. Moreover, inherent biases in training data could perpetuate unfair practices, and the lack of legal authority of an AI summary compared to the original document must be transparently communicated. Future development must prioritize rigorous validation, clear disclaimers, explainable AI techniques to build trust, and continuous updates, all informed by interdisciplinary collaboration with legal and ethics experts to ensure responsible and genuinely empowering AI solutions.

References

[1] W. Ahmad, J. Chi, Y. Tian, and K.-W. Chang. Policyqa: A reading comprehension dataset for privacy policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.66

[2] J. Chi, W. U. Ahmad, Y. Tian, and K.-W. Chang. 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)*. 2023. https://doi.org/10.18653/v1/2023.acl-short.31

[3] C. M. Greco and A. Tagarelli. Bringing order into the realm of transformer-based language models for Artificial Intelligence and law. *Artificial Intelligence and Law*, 32(4):863–1010, 2023. https://doi.org/10.1007/s10506-023-09374-7

[4] M. R. Parvez, J. Chi, W. U. Ahmad, Y. Tian, and K.-W. Chang. Retrieval enhanced data augmentation for question answering on privacy policies. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 201–210, 2023. https://doi.org/10.18653/v1/2023.eacl-main.16

[5] D. Song, S. Gao, B. He, and F. Schilder. On the effectiveness of pre-trained language models

for Legal Natural Language Processing: An empirical study. *IEEE Access*, 10:75835–75858, 2022. https://doi.org/10.1109/access.2022.3190408

[6] B. Auxier. Americans and privacy: Concerned, confused and feeling lack of control over their personal information. *Pew Research Center*, November 15, 2019. https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information

[7] Chi, Jianfeng and Ahmad, Wasi Uddin and Tian, Yuan and Chang, Kai-Wei. PLUE: Language Understanding Evaluation Benchmark for Privacy Policies in English, https://github.com/JFChi/PLUE