ISSN 1870-4069

Self-Supervised Learning with Legal-Related Corpus: Customizing a Language Model with Synthetic Data

Philippe Prince-Tritto[1], Hiram Ponce[2]

[1] Universidad Panamericana,
Facultad de Derecho,
Mexico

[2] Universidad Panamericana,
Facultad de Ingeniería,
Mexico

{pprince, hponce}@up.edu.mx

Abstract. This paper explores the development of a customized text generation system using pre-trained language models, specifically aimed at knowledge workers such as lawyers and personal data protection specialists. Our approach minimizes human intervention in the labeling process for fine-tuning. To this end, we automate data collection and filter the data through GPT-3.5 and BERT-based heuristics. Human expertise is only leveraged in design and oversight, ensuring the system's ability to provide accurate and relevant information. We also developed an annotation tool to complete our training set, utilizing text generation which required a low level of human supervision. This paper repurposes the Prompt Generation Network architecture to create a chatbot in Spanish language that can address queries related to personal data protection. Our results showed encouraging progress towards automating the annotation of a dataset for fine-tuning with little human intervention, although opportunities for improvement remain. Ultimately, our research offers a blueprint for the creation of a chatbot using a fine-tuned language model with minimal human intervention, demonstrating the potential of these models for practical applications.

Keywords: Self-supervised learning, legal language processing, fine-tuning, large language model.

1 Introduction

The rapid advancement in the field of Machine Learning (ML) and natural language processing (NLP) has led to the development of increasingly sophisticated language models, capable of analyzing and predicting human-like text. These models, such as GPT [9] and LLaMa [10], hold immense potential for a wide range of applications, from personalized assistance to professional tools for knowledge workers.

However, harnessing the power of these pre-trained models often requires a fine-tuning process that can be time-consuming and resource-intensive, particularly when it comes to the labeling of training data. In this paper, we explore the possibility of utilizing pre-trained language models to create a customized text generation system that caters to the specific needs of knowledge workers, such as lawyers and specialists in personal data protection. Our objective is to minimize human intervention by automating the labeling process using pre-trained models like GPT-3.5 and BERT for data filtering and question identification. Human expertise is reserved only for design and oversight tasks that require nuanced judgment. Upon existing research on the efficiency of pre-trained language models, prompt-learning architectures, and the scalability of text generation from prompts, we can claim that these techniques can be combined to develop a robust, lightly supervised annotation tool.

Some authors have focused primarily on the effectiveness of pre-trained language models to be refined to recognize and respond to specific prompts [11]. The proposed architectures are intended to be applied in the pre-training phase to improve the efficiency of fine-tuning. For this paper, we are inspired by the Prompt Generation Network (PGN) architecture [6] for Prompt-learning.

PGN consists in generating input-dependent prompts by sampling from a learned library of tokens. It should be noted that for these authors, the task-specific data are in the pixel space, while in our case it is text. The scalability of text generation from prompts could provide immense potential for the development of robust annotation tools that require fairly low human supervision. Transformers systems have the ability to generate their own tags and patterns for learning how to learn.

They are able to do this in a way that is more efficient and less costly than human-generated annotation [12]. This can also alleviate the shortage of labeled data, which is an obstacle to better performance, and thus enrich the model representations [5]. It should be understood that today, the main bottleneck when training a language model, or any supervised learning, is data labeling. However, some [3] have shown that the use of pre-trained models can surpass human intervention in this repetitive, mind-numbing task for the operator, and with little added value from a humanistic perspective.

In this study, the role of human intervention was primarily in the design and oversight of the data collection, with a thorough ETL process, and model fine-tuning processes. The web scraping process was automated but designed by human engineers. The initial dataset of articles was filtered using a GPT-3.5 model, reducing the need for human curation. However, human judgment was applied in the design of the BERT-based heuristic for question identification and in the choice of hyperparameters for model fine-tuning. The aim was to minimize human involvement in the routine tasks of data labeling and curation, while still leveraging human expertise for tasks that required nuanced judgment.

We aim to demonstrate that the availability of language models such as GPT or LLaMa opens up a potential for customizing ML models for knowledge workers. The problem is thus the following: How can we use pre-trained language models to produce a text generation system that relies on specific knowledge, with the least amount of human intervention with respect to the labeling of the model's fine tuning data?

To answer this, we developed a Spanish chatbot capable of addressing laypeople's queries related to personal data protection. Fine tuning a specific LLM requires data labelized as prompt and completions. For the completions, we implemented our approach by scraping specific web databases comprising privacy-related newspaper articles, cleaning the data, and separating it into target responses.

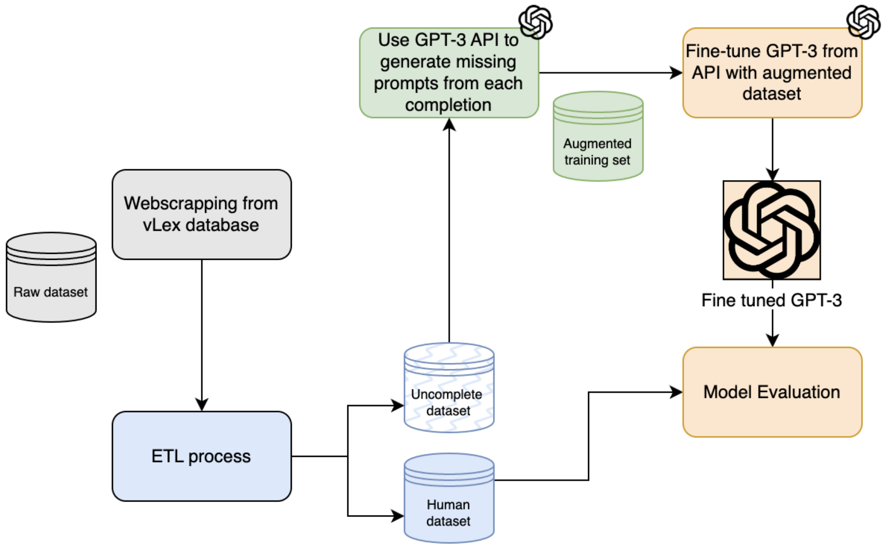

Fig. 1. Flowchart diagram for the self-supervised proposal.

These responses needed questions: the corresponding 'prompt'. GPT-3.5 was utilized to generate the prompt for each given completion for fine-tuning the LLM without expert intervention. The evaluation involved (1) a fully human-curated dataset, derived from the web scraping step, and (2) a partially synthetic dataset, where prompts were generated by GPT-3.5 and completion was taken from the scraped data.

Each dataset provided a basis for comparing the responses generated by the fine-tuned model to the target output. The rest of the paper is organized as follows. In Section 2, we present the working experiment of creating a Spanish-language chatbot capable of addressing questions and concerns related to personal data protection, from an non-annotated dataset.

We detail the experimentation's execution, and then examine the efficacy of this chatbot in Section 3. After demonstrating the potential for customized ML models to assist knowledge workers in their professional endeavors, we conclude in Section 4 on the use of the advancements in NLP without placing undue burden on human resources.

2 Materials and Methods

We propose to fine-tune a large language model to create an agent for lay people to resolve their doubts about the requirements of personal data protection in Spanish language and in Mexico. The focus of the experiment is on the dataset used for the model fine tuning that underlies the chatbot. We exploit a pre-trained model to generate sound training data. Following [7], we go one step further incorporating some synthetic data in the training database to improve its performance, as shown in Figure 1.

The database consists of non-annotated legal news articles from vLex, obtained with webscrapping techniques. We assume that the authors of these news articles are answering a specific question about personal data. What is missing is a formulation of the question they answer. It is this question, the prompt, that we propose to generate with a pre-trained language model.

We then reuse this synthetic data to fine-tune this same language model. Our agent is a chatbot that generates through it's interface text fragments on the specific topic of personal data protection in Spanish. The experiment takes the form of a proof-of-concept to validate the hypothesis.

2.1 ETL Process

Description of the Initial Dataset. The first task is to retrieve data related to privacy. In order to work with real data and replicate the process used in a company wishing to exploit its own data, we downloaded via a webscrapping process the data from the vLex platform, searching for the exact term "Protección de datos personales" (Personal data protection) and filtering by type of documents.

To carry out our experiment, we limited ourselves to press articles on the subject. As the articles were classified by relevance, we used this classification for webscrapping and retrieved the first 4091 results. To remove noise from the retrieved data, we filtered the results using a binary classification with GPT-3.5, based on the title of the article. The prompt used for the classification task was:

```
  Olvida todas las instrucciones anteriores. Eres un
clasificador de noticias en materia de privacidad en México.
Basándose en el análisis del título, tu tarea consiste en
responder '1' si es probable que el artículo trate de la
protección de datos personales, y '0' en caso contrario. El
título es: 'title'.
```

The English translation of the prompt is:

```
  Forget all previous instructions. You are a news classifier
focused on privacy matters in Mexico. Based on the title
analysis, your task is to answer '1' if it's likely that the
article is about personal data protection, and '0' otherwise.
The title is: '{title}'.
```

This allowed us to determine that the number of articles directly related to privacy was 737. This simple step using a heuristic based on an existing language model allowed us to keep only relevant data at minimal cost. It seemed reasonable to not webscrap the whole vLex database, since we did not have a lot of relevant items after reaching a certain point. We determined this point with a plot that shows a smoothed rate of change in the number of privacy related articles.

When the smoothed rate of change approaches zero, it indicates that there is no significant increase in the number of privacy-related documents anymore, as shown in Figure 2. The articles were then separated into paragraphs, with each paragraph representing a human-data sample.

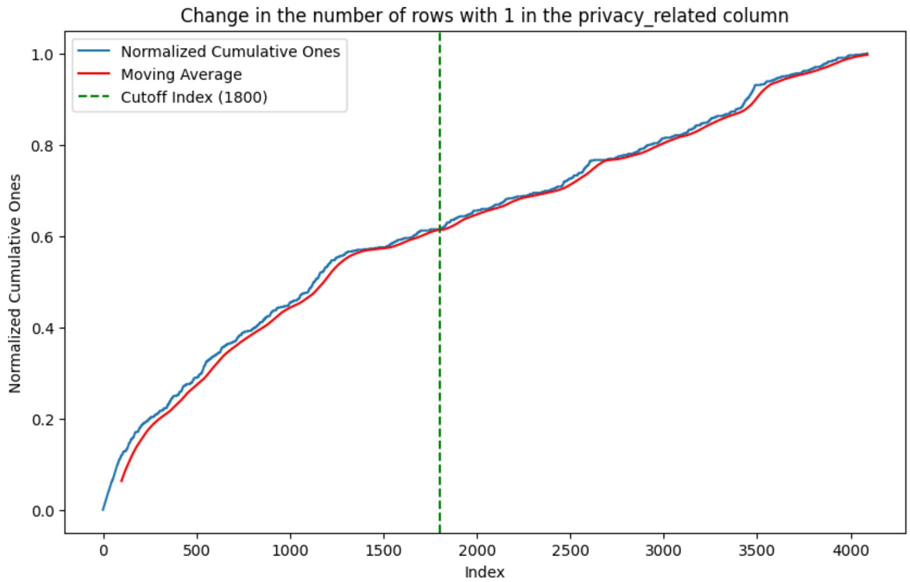

Fig. 2. Smooth rate of change in the privacy-related document. The cutoff index is percentile-based so that the threshold is equal to 0.00003.

Each sample was normalized, removing the logical connectors at the beginning of the paragraph to make it look like an answer to a question. As many irrelevant paragraphs as possible were removed from the dataset, for example when they began with certain cues that announced they were advertising paragraphs. We also utilized spaCy to eliminate samples that began with a proper name, as this information was not pertinent to a general chatbot focusing on personal data.

Most paragraphs of this nature lacked substance but instead provided details such as the author's identity and a summary of their professional background. To build a base of Full Human data, we then identified the questions in our dataset, storing them as 'prompt'. The following paragraphs were stored as 'completion'. In order to determine which paragraphs could be used as answers, we performed a Similarity-based heuristic with BERT.

Due to the small number of questions identified in our dataset, we also assumed that the shortest paragraphs were titles and we based our heuristic on the fact that the first paragraphs following a title give a short answer to it. We therefore stored as 'prompt' headlines of 15 words or and concatenated "What can you tell me about" with the headline. We then repeated the Similarity-based heuristic with BERT to store the following paragraphs as 'completion'.

From the 737 privacy related articles, the above method allowed us to get to 4177 human-data samples that constituted the completions to the synthetic prompts we would further generate, and 545 tuples of Full Human-data prompts and completions, as shown in Figure 3. Full Human dataset would further be used for testing purposes only. After estimating the cost of fine-tuning an OpenAI davinci model, including the generation of synthetic prompts, we moved on to generating the training set with synthetic data.

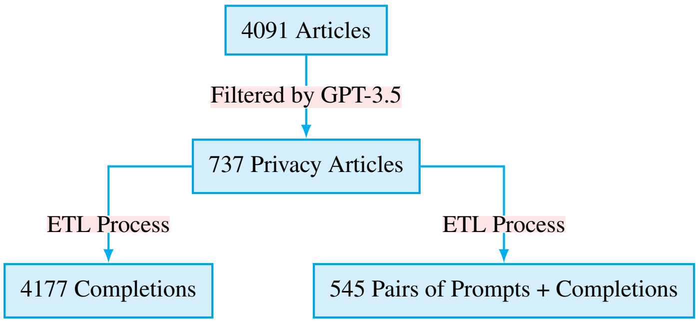

Fig. 3. From webscraping to filtered dataset for synthetic data generation for fine-tuning and full human validation set.

Prompt Generation for Synthetic Data and Final Dataset Summary On a sample of 10 completions, we performed a grid search to determine the hyperparameters of the davinci model of GPT-3.5 as shown in Table 1. As for the Partially synthetic training and testing data, we designed a prompt to guide the language model in its response. It goes as follows:

```
  Eres experto en protección de datos personales en México.
Genera un prompt conciso y corto en idioma Español que podría
ser el mejor candidato para ser contestado por el siguiente
texto en materia de protección de datos personales en México:
``{row[``completion'']}''
```

The English translation of the prompt is:

```
  You are an expert in personal data protection in Mexico.
Generate a concise and short prompt in Spanish language that
could be the best candidate to be answered by the following
text on the subject of personal data protection in Mexico:
``{row[``completion'']}''
```

In other words, the Partially Synthetic dataset is composed of tuples of synthetic-prompt and human-completion samples. Generating the prompts corresponding to each of the completions resulted in a dataset of 4177 tuples of synthetic-prompt and human-completion samples. This dataset was separated such that 16% constituted the Partially Synthetic test set for the fine tuning of the model, and the remainder the Partially Synthetic training set. Some examples are given in Table 2.

As described above (see Figure 3), the Full Human test set to validate the fine-tuned model is composed of 545 data points. This dataset is composed of tuples of human-prompt and human-completion samples. Some examples are given in Table 3. We represented the datasets summary in Table 4. An essential aspect of our methodology is the validation of the generated data. While the synthetic prompts are generated by a fine-tuned language model, their pairing with human-generated completions ensures a level of quality and relevance.

Table 1. Hyperparameter Grid search for Prompt Generation (Results in bold).

Temperature	Top-P	Frequency Penalty	Presence Penalty
0.5	0.5	0.5	**0.5**
0.8	**0.8**	0.8	0.8
1.0	1.0	**1.0**	1.0

These pairs undergo a systematic filtering process, as outlined in Section 2.1, to remove any outliers or irrelevant entries. Similarly, the Full Human dataset is derived from vetted, privacy-related articles, adding another layer of quality control. No manual corrections are applied to the data; instead, we rely on the rigor of our automated processes and the fine-tuning and evaluation metrics to ensure data integrity.

We further used the synthetic-prompt from the Partially Synthetic test set and human-prompt from the Full Human test set to further validate the fine-tuned model, generating completion from those prompts and comparing it with the human-completion of the Partially Synthetic test set and the Full Human test set.

Model Fine-Tuning. For the fine-tuning process, we exclusively utilized the Partially Synthetic dataset. This contextual detail is pivotal for interpreting the subsequent performance evaluation of the fine-tuned model.

We used Weight and Bias for monitoring and the OpenAI API to fine tune the davinci GPT model. The hyperparameters chosen for this step were not subject to a grid search because of the cost that this could represent. For the hyperparameters, we used a Batch size of 64, which is the high limit of what is commonly practiced, a learning rate of 0.01 and 4 epochs, in order to avoid overfitting while preserving training costs.

Finally, since all tasks are equally important in the task of our language model, we set a prompt loss weight of 1.0. The OpenAI API for fine-tuning allowed us to measure the loss (for assessing if the model is learning and fitting the training data well and performs well with unseen examples) and token accuracy (for assessing if the model predicts the correct token) for both the training and validation sets.

It is worth noting that when fine-tuning the model, all layers are retrained, since fine-tuning is a process that adjusts all the weights and biases in the model, across all layers [4]. The purpose of fine-tuning is to adapt a pre-trained model, which was originally trained on a large, diverse dataset, to perform well on a specific task or to better match a narrower dataset.

2.2 Performance Evaluation

To evaluate the model performance, we use two validation sets of questions that the model has not seen before. We compare the agent's responses to the correct answers. Two tests are performed. The first one, where both the prompt and the answer are natural data (Human Set Validation, see Table 6). The other one, where the prompt is synthetic data and the answer is natural data (Synthetic Set Validation, see Table 6). We then proceed both to a quantitative and a qualitative assessment.

Quantitative Assessment. We are unable to calculate perplexity for the generated texts, as this would require knowing the exact probabilities that the model assigns to each word, which is not available with GPT-3.5.

Table 2. Example of synthetic prompts and human completions from the Partially Synthetic Dataset.

Prompt	Completion
¿Cómo se puede prevenir el trashing en México?	Este delito se conoce como trashing y consiste en que los delincuentes obtienen información privada como estados de cuenta, copias de identificaciones oficiales, recibos, documentos, directorios e incluso contraseñas que el usuario ha enviado a la basura o a la papelera de reciclaje de su equipo. Existe tanto de manera física como digital.
¿Qué implicaciones tiene el RFC para la protección de datos personales en México?	La idea de que todos los mayores de 18 años en México cuenten con su Registro Federal de Contribuyentes (RFC) es mucho más que la simple búsqueda de un control fiscal de las autoridades tributarias sobre los ciudadanos. Obligar a todos a tener un RFC forma parte de los esfuerzos del Gobierno federal de construir una matrix para la vigilancia masiva e indiscriminada, en línea con el registro digital de ciudadanos que impulsa la Secretaría de Gobernación y de la construcción de un padrón nacional de datos biométricos de usuarios de telefonía móvil.
¿Qué medidas propone Lorenzo Córdova para garantizar la protección de datos personales en México?	En tanto, Lorenzo Córdova se mostró en la disposición de avanzar en este tema, siempre y cuando no se vulnere el derecho a la protección de datos personales de los ciudadanos, así como que no se viole ninguna normativa que rija a la Secretaría de Gobernación y al Instituto Nacional Electoral.
¿Cuáles son las consecuencias de la omisión cometida por Banamex en materia de protección de datos personales?	De acuerdo con el IFAI, Banamex contravino los principios de consentimiento de calidad y responsabilidad; mantuvo datos inexactos del titular y no hizo la rectificación o cancelación de los mismos, no obstante que resultaba legalmente procedente, por lo que se configuró una omisión que hizo nugatorio el derecho ejercido por el particular.

However, we need to assess the LLM's responses. To this end, it is necessary to determine if the output is correct. We do that using simple unique word overlap, tokenizing both the target and the generated text into individual words and then comparing these sets of words. We calculate similarity scores but do not establish a threshold for determining if a generated text is correct or not.

In other words, True Positives (TP) are the words in the generated text that are also in the target text, False Positives (FP) are the words in the generated text that are not in the target text, and False Negatives (FN) are the words in the target text that are not in the generated text. From there, we were able to compute set-based Precision, Recall and F1 score, as well as ROUGE scores:

– **Precision.** is computed as in Equation 1:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{1}$$

Table 3. Example of human prompts and completions from the Full Human Dataset.

Prompt	Completion
¿Qué información no comparte WhatsApp con Facebook?	En la política de privacidad, WhatsApp asegura que limita significativamente la información que comparte con Facebook. WhatsApp señala que no comparte: No mantiene registros de con quién hablas o envías mensajes.
¿Qué me puedes decir sobre ‘Drones: Opinion 01/2015 on Privacy and Data Protection Issues relating to the Utilisation of Drones’?	Los esfuerzos por garantizar el derecho a la protección de datos personales tienen la finalidad de otorgar a los usuarios de los medios digitales mayor control sobre el uso de su información en manos de organizaciones, sujetos obligados y particulares. En la actualidad, la protección de datos en nuestro país cuenta con un marco jurídico sólido para garantizar su ejercicio.
¿Cuál es el papel de los ciudadanos en la protección de sus datos?	Deben estar conscientes de que su información personal vale y que ante las nuevas tecnologías de la información debemos tener mucho cuidado con los datos que proporcionamos, con las medidas de seguridad de nuestros dispositivos electrónicos. En el uso de la redes sociales es fundamental revisar las declaraciones y los avisos de privacidad.
¿Qué me puedes decir sobre ‘Aviso de privacidad’?	El aviso de privacidad garantiza la salvaguarda del derecho de autodeterminación informativa reconocido en las normatividades de protección de datos personales en México.

– **Recall:** Also known as Sensitivity, is computed as in Equation 2:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{2}$$

– **F1 score:** As shown in Equation 3, is the weighted average of Precision and Recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

– **ROUGE score:** (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics[1] for evaluating automatic summarizing of texts as well as machine translation. ROUGE-1 and ROUGE-2 are computed as in Equations 4, 5 and 6, considering the overlap of 1-grams and 2-grams.

ROUGE accounts for the frequency of each word, meaning that duplicate words in both the generated text and the target text are considered. This lead to different Precision and Recall values compared to the aforementioned set-based approach, so it can give a broader scope for evaluation:

Table 4. Datasets used in the self-supervised learning experiment.

Dataset	Train	Test
Partially Synthetic	3480	697
Full Human	0	545

$$\text{ROUGE-N Precision} = \frac{\text{Number of overlapping N-grams}}{\text{Total N-grams in the generated text}}, \tag{4}$$

$$\text{ROUGE-N Recall} = \frac{\text{Number of overlapping N-grams}}{\text{Total N-grams in the target text}}, \tag{5}$$

$$\text{ROUGE-N F1 Score} = 2 \times \frac{\text{ROUGE-N Precision} \times \text{ROUGE-N Recall}}{\text{ROUGE-N Precision} + \text{ROUGE-N Recall}}, \tag{6}$$

where N is the length of the n-gram (e.g., for ROUGE-1, $N = 1$ and the n-grams are individual words; for ROUGE-2, $N = 2$ and the n-grams are two consecutive words, etc.). On the other hand, ROUGE-L considers the Longest Common Subsequence (LCS) between the generated and target texts as shown in Equations 7, 8 and 9. The LCS is a sequence of words that appear in the same order in both texts, although not necessarily consecutively:

$$\text{ROUGE-L Precision} = \frac{\text{Length of LCS}}{\text{Total number of words in the generated text}}, \tag{7}$$

$$\text{ROUGE-L Recall} = \frac{\text{Length of LCS}}{\text{Total number of words in the target text}}, \tag{8}$$

$$\text{ROUGE-L F1 Score} = 2 \times \frac{\text{ROUGE-L Precision} \times \text{ROUGE-L Recall}}{\text{ROUGE-L Precision} + \text{ROUGE-L Recall}}. \tag{9}$$

We calculated ROUGE scores with the 'rouge.get_scores()' function from the 'rouge' Python package. The function returns the F1 scores directly for ROUGE-1, ROUGE-2, and ROUGE-L. However, this function calculates these metrics internally in a similar manner to the formulas mentioned above.

These metrics collectively offer a multi-faceted view of how well the generated text matches the target text, although they do not measure the correctness of the generated text in a semantic or qualitative sense. We also computed BLEU and METEOR scores (overlap between the generated summary and a reference summary) to further validate the answers.

– **BLEU:** (Bilingual Evaluation Understudy) score [8] is a metric that has been developed to evaluate machine translation systems. It compares the n-grams in the generated output and the target, and gives scores between 0 and 1, where 1 is the perfect score as in Equation 10:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log(p_n)\right), \tag{10}$$

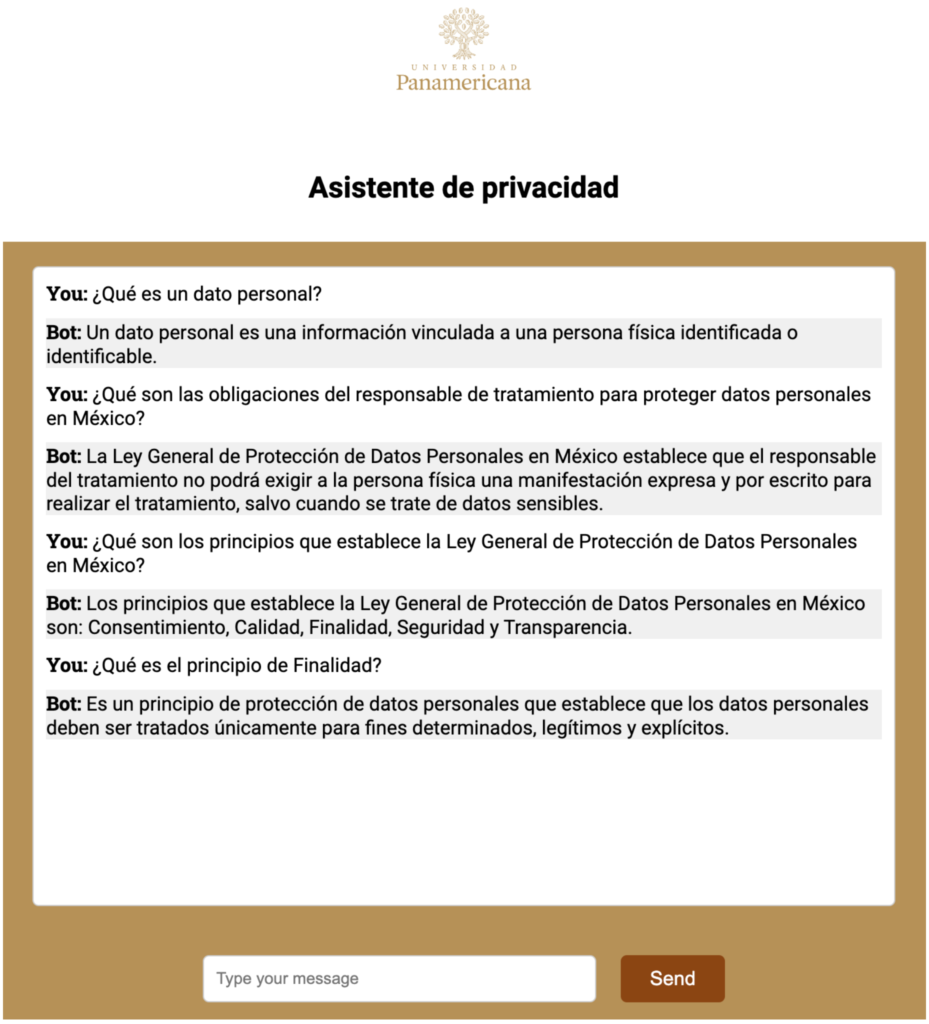

Fig. 4. Chatbot interface showing some prompts and responses.

where:
p_n = is the precision for n-grams.
w_n = is the weight for each n-gram with $w_n = 1/N$.
N = is the maximum order of n-grams used.
BP = is the brevity penalty, calculated as:

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r, \\ \exp\left(1 - \dfrac{r}{c}\right) & \text{if } c \le r, \end{cases} \tag{11}$$

where c is the length of the candidate translation and r is the effective reference corpus length.

– **METEOR:** (Metric for Evaluation of Translation with Explicit ORdering) is a metric[2] that has been developed to overcome some of the limitations of metrics such as BLEU. It gives scores between 0 and 1, where 1 is the perfect score. The overall METEOR score is then calculated as in Equation 12:

$$\text{Score} = (1 - \text{Penalty}) \cdot \text{Fmean}, \tag{12}$$

where Penalty is calculated based on the number of chunks (c) and total number of matched unigrams (m) as Penalty $= 0.5 \cdot (\text{c/m})^3$, and Fmean is the harmonic mean of Precision (P) and Recall (R), with a parameter α set to 0.9 to weight recall more heavily so that:

$$\text{Fmean} = \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R}. \tag{13}$$

Qualitative Assessment. To carry out a qualitative evaluation of the large language model's output, a human evaluator expert in data privacy rated the generated texts based on a defined set of criteria designed to capture important aspects of text quality that are relevant to the evaluation of text generation systems. We assigned weights to each criterion based on their relative importance. The weights w_n reflect the priorities and preferences of the evaluation process.

The evaluation criteria includes: Relevance ($w_1 = 0.2$), which assesses if the generated text aligns with the topic of personal data protection; accuracy ($w_2 = 0.3$), which scrutinizes the correctness and up-to-date nature of the information in the generated text; understandability ($w_3 = 0.15$), which examines if the generated text is easily comprehensible by the target audience; completeness ($w_4 = 0.2$), which measures if the generated text covers all the relevant aspects of the subject; objectivity ($w_5 = 0.1$), which checks for the impartiality and balanced presentation of information; and structure and coherence ($w_6 = 0.05$), which evaluates if the generated text is logically consistent and well-structured.

For each criterion, we compute the average rating from the evaluators, ranging from 0 to 100. We denoted these average ratings as (r_1), (r_2), (r_3), (r_4), (r_5), and (r_6), respectively. We then calculate the Qualitative Score across all evaluators for each criterion, considering the weighted importance, as shown in Equation 14:

$$\text{Qualitative Score} = \frac{\sum_{i=1}^{6} w_i \cdot r_i}{\sum_{i=1}^{6} w_i}. \tag{14}$$

The qualitative assessment was performed on 100 random samples: 50 from the Human Set Validation and 50 from the Synthetic Set Validation.

2.3 Chatbot Implementation

The chatbot implementation makes use of the Flask web framework and the OpenAI's GPT models, leveraging the OpenAI API for conversational responses.

Table 5. Fine Tuning Results.

Metric	Value
Training Loss	0.469
Training Token Accuracy	0.682
Validation Loss	0.729
Validation Token Accuracy	0.676

Two routes are defined - '/' and '/chat'. The first route displays the HTML front end for the chatbot while the second route performs the chatbot processing.

```
@app.route('/')
def index():
return render_template('07_chatbot_front.html')

@app.route('/chat', methods=['POST'])
def chat():
user_message = request.json['message']
```

When the /chat route is accessed, it retrieves the user's message from the JSON payload of the POST request. This message is then used as a prompt to generate a completion from the OpenAI model, with the following hyperparameters:

```
max_tokens=200,
temperature=0.5,
top_p=0.8,
frequency_penalty=1.0,
presence_penalty=0.5,
stop=[``\n'']
```

After receiving a completion, it is processed to remove leading and trailing white space and replacing occurrences of "-¿" (markdown formatting used by our fine-tuned model for answering). The processed completion is then returned to the user interface as a JSON payload, and ploted in an HTML page as shown in Figure 4. The chatbot can be viewed by running the Flask application and opening the specified URL in a web browser, typically 'localhost' with the assigned port number '5000'.

3 Results and Discussion

In this section, we present the results of our experiment, which aimed to fine-tune a large language model for answering personal data protection questions in Spanish language and Mexico. We also discuss the implications of these results in terms of the performance of the chatbot agent.

3.1 Results

The overall results for our fine-tuned model are presented in Table 5. The performance of the chatbot agent was first evaluated using a set of validation questions not seen during training.

The agent's responses were compared to the correct answers, and various evaluation metrics were computed, including precision, recall, F1 score, BLEU score, ROUGE scores and METEOR score. The results are summarized in Table 6. We then carried out the manual human evaluation for qualitative scoring. The results are summarized in Table 7. We also found that the length of generated texts was on average 323% greater when human prompts were presented (178 words) to the fine-tuned model, than when synthetically generated prompts were presented (42 words).

3.2 Discussion

Presenting these separate evaluations in Tables 6 and 7 between the Fill Human and Partially Synthetic sets highlights how well the fine-tuned model performs in different settings—responding to human-generated prompts and synthetic prompts.

We recognize that the Human and Synthetic sets are inherently different, and this is explicitly by design. The Partially Synthetic set is created for the primary purpose of fine-tuning, while the Full Human set serves as a more naturalistic ground truth for performance validation. Therefore this comparison is not intended to show that one is better than the other.

Instead, it offers a multi-faceted evaluation of the model's capabilities. These separate evaluations provide a comprehensive understanding of the model's performance. Based on the metrics presented in Table 5, the model appears to be performing reasonably well, with relatively low training loss (0.469 and 0.729) and moderate token accuracy (0.682 and 0.676) on both training and validation datasets.

However, it's important to consider the specific requirements of personal data protection contents generation. The results presented in Table 6 could be interpreted as a relatively low performance of the chatbot agent in answering questions about personal data protection, suggesting that there is considerable room for improvement in the chatbot's ability to accurately answer questions on this topic. The low ROUGE scores, especially the ROUGE-2 score, also indicate that the generated answers do not closely match the reference summaries.

This could be due to several factors, such as the quality of the training dataset or the limitations of the fine-tuning process. Additionally, the synthetic data generation process may have introduced noise or biases into the training data, which could have negatively impacted the performance of the chatbot. However, it should be noted that for creative or low constrained tasks, such as text generation, it's difficult to assess the quality of an output via a quantitative metric.

Precision, Recall, F1, ROUGE, BLEU and METEOR scores may work as general indicators, but could not be very informative. In these cases, good output can vary enormously, and output that doesn't exactly match the target can still be considered good quality. In addition, these scores are similarity measures based on the presence of common unigrams, bigrams, etc., in the generated output and the target.

They do not capture the semantics or meaning of the output. For example, an output that uses synonyms of words in the target might be semantically very similar to the target, but would have a low ROUGE, BLEU or METEOR score. An important aspect of our findings resides in the analysis of the textual output derived from human-generated prompts and synthetic-generated prompts.

Table 6. Performance evaluation metrics of completions for the fine-tuned model.

Metric	Human Set	Synthetic Set
Average BLEU score	0.0028	**0.0236**
Average Precision	0.1792	**0.3425**
Average Recall	0.2053	**0.3425**
Average F1 Score	0.1521	**0.3350**
Average ROUGE-1 F-score	0.1179	**0.2045**
Average ROUGE-2 F-score	0.0151	**0.0421**
Average ROUGE-L F-score	0.0949	**0.1506**
Average ROUGE-1 Precision	0.1597	**0.2124**
Average ROUGE-2 Precision	0.0220	**0.0444**
Average ROUGE-L Precision	0.1308	**0.1570**
Average ROUGE-1 Recall	0.1431	**0.2124**
Average ROUGE-2 Recall	0.0206	**0.0449**
Average ROUGE-L Recall	0.1167	**0.1564**
Average METEOR score	0.1028	**0.1797**

Our findings reveal a noteworthy trend: the model performs significantly better on synthetic prompts compared to human-generated prompts across multiple evaluation metrics. While it may be tempting to attribute this solely to the model being fine-tuned on synthetic data, it is essential to recognize that these results offer valuable insights into the general interplay between synthetic and human-generated data in natural language processing tasks.

The superior performance with synthetic prompts illuminates possible advantages in their structural and stylistic attributes that make them more conducive for machine interpretation and response generation. This highlights a broader question about the efficacy and limitations of machine learning models in simulating human-like conversational abilities.

It also raises the issue of whether the model's current configuration is sufficiently robust to handle the nuances and complexities inherent in human language. These insights serve to enrich the ongoing discourse on the balance between training data types and model performance, and provide a compelling avenue for future research.

In terms of the qualitative assessment, as presented in Table 7, the generated outputs from synthetic prompts were scored higher in all evaluation criteria, with the overall Quality Score being 55.1 compared to 42.9 for responses generated from human prompts. This suggests that the model performed better when dealing with prompts generated synthetically, indicating a successful transfer of learning.

However, it was noted that the scores for completeness were lower for outputs generated from synthetic prompts, maybe due to the human prompt structure. In addition, the length of generated texts was much longer when human prompts were used, suggesting that synthetic prompts likely lead to more concise responses.

Table 7. Manual human evaluation of completions for qualitative scoring.

	Relevance	Accuracy	Understandability	Completeness	Objectivity	Structure and Coherence	Quality Score
Generated from human prompt	53.0	29.2	59.5	39.1	37.8	**61.3**	42.9
Generated from synthetic prompt	**75.8**	**44.8**	**86.3**	**32.0**	**51.0**	52.9	**55.7**

Further work could investigate if this pattern holds for different domains or languages. There exist several potential avenues for future research aimed at enhancing the efficacy of the chatbot agent. One avenue involves the refinement of the synthetic data generation procedure to produce prompts of superior quality for the training dataset. Another approach entails the inclusion of supplementary sources of training data, coupled with continued efforts to augment the quality and quantity of the training data through the ETL process.

It is important to acknowledge that our study utilized a dataset of relatively modest size, and thus, efforts should be made to enhance its size and diversity, particularly considering that solely news articles were employed for a chatbot that had a legal-related task. Moreover, the qualitative evaluation could be enriched by engaging multiple experts, thereby facilitating a more comprehensive and unbiased assessment. Lastly, further investigation into distinct fine-tuning strategies, hyperparameter optimization, and model architectures has the potential to yield advancements in the chatbot's performance.

4 Conclusions

The chatbot agent demonstrated a limited ability to accurately answer questions about personal data protection in the Spanish language and in Mexico. However, the proposed method for fine-tuning a large language model, specifically for answering personal data protection questions in Spanish, yielded encouraging results. It confirms that a combination of real and synthetic data for fine-tuning can indeed lead to coherent generation of domain-specific text.

In this work, we validated that automation of the annotation of a dataset for fine-tuning is possible with minimal human intervention, primarily focused on design and oversight tasks. It would be necessary to repeat the experiment with a more substantial and diverse dataset, not just legal journalism data, and a larger number of epochs. Our method provides a blueprint for the creation of a chatbot by using fine-tuned language models with few human intervention.

We underlined the potential of these models in practical applications, such as a data protection chatbot that can provide understandable and accurate information to lay users. On a broader scale, this experimental approach highlights how machine learning models can be adapted to specific tasks or domains with the help of fine-tuning strategies, even when a substantial amount of specific task-related training data is not available. It also provides insights for future research on the specifics of fine-tuning these models, which will be increasingly relevant as applications of large language models continue to expand.

One limitation inherent to our methodology was the exclusive utilization of synthetic prompts for fine-tuning the model. Future research endeavors could potentially employ a balanced blend of Human and Synthetic prompts for fine-tuning to engender a model with more robust generalizability across different data domains.

For future work, we will also consider improving the method by integrating a more comprehensive dataset that includes more diversity in terms of topics, formats, and writing styles. In particular, incorporating legal texts, regulatory guidelines, and court case summaries related to personal data protection could enhance the model's understanding of this specific field.

As for evaluation, we could employ a more robust qualitative assessment, involving a larger panel of domain experts, to better gauge the semantic quality and relevance of the chatbot responses. Another important direction of research is the exploration of causal AI techniques to improve the quality of the responses it generates. Causal AI is an area of machine learning that builds models based on causal relationships rather than mere correlations.

This approach could be particularly useful in legal contexts, such as personal data protection, where understanding the cause-and-effect relationships between different elements of the law is crucial. One potential avenue to explore is the use of causal inference techniques to understand which elements of the training data have the most significant impact on the chatbot's performance.

By identifying these causal relationships, we could optimize the training process and focus on the most influential data elements. Further, integrating counterfactual reasoning within the chatbot may prove beneficial. Counterfactual reasoning is a core component of causal AI, enabling the model to consider alternate scenarios and outcomes, an ability particularly relevant in a legal context. For instance, understanding how a different data protection regulation could affect a certain scenario could be invaluable for users.

Acknowledgments. We would like to acknowledge the assistance provided by ChatGPT, developed by OpenAI. ChatGPT was used to enhance the syntax and wording of this paper, particularly in the editing and refinement of the manuscript. While ChatGPT offered suggestions and improvements, the responsibility for the scientific content and ideas presented in this paper remains solely with the authors.

References

1. Chin-Yew, L.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. Association for Computational Linguistics, pp. 74–81 (2004)
2. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376–380 (2014)
3. Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT outperforms crowd workers for text-annotation tasks. In: Proceedings of the National Academy of Sciences, vol. 120 (2023) doi: 10.1073/pnas.2305016120

4. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 328–339 (2018) doi: 10.48550/arXiv.1801.06146
5. Liu, B., Lin, T., Li, M.: Enhancing aspect-category sentiment analysis via syntactic data augmentation and knowledge enhancement. Knowledge-Based Systems, vol. 264, pp. 110339 (2023) doi: 10.1016/j.knosys.2023.110339
6. Loedeman, J., Stol, M. C., Han, T., Asano, Y. M.: Prompt generation networks for efficient adaptation of frozen vision transformers. In: Proceedings of the International Conference on Learning Representations (2022) doi: 10.48550/arXiv.2210.06466
7. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Proceedings of the 36th Conference on Neural Information Processing Systems. Advances in Neural Information Processing Systems, vol. 35, pp. 27730–27744 (2022)
8. Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002) doi: 10.3115/1073083.1073135
9. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
10. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and efficient foundation language models (2023) doi: 10.48550/arXiv.2302.13971
11. Wang, J., Wang, C., Luo, F., Tan, C., Qiu, M., Yang, F., Shi, Q., Huang, S., Gao, M.: Towards unified prompt tuning for few-shot text classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2022) doi: 10.48550/arXiv.2205.05313
12. Watson, E., Viana, T., Zhang, S.: Augmented behavioral annotation tools, with application to multimodal datasets and models: A systematic review. AI Multidisciplinary Digital Publishing Institute, vol. 4, no. 1, pp. 128–171 (2023) doi: 10.3390/ai4010007