Anonimización de Documentos Legales usando LLMs con Preentrenamiento Continuo y Sintonizado Fino

Sofia Ornella Ortman*[1,2], Luciana Belen Canteros*[1,2], Francisco Vargas[2], Gaston Escalante[2], Alejandro González Coene[2], Manuel Pulido[1,3]

[1] FaCENA, Universidad Nacional del Nordeste, Corrientes
[2] Legalhub S. A., Buenos Aires
[3] Instituto de Modelado e Innovación Tecnológica, CONICET, Argentina
ortmansofia00@gmail.com, 2010lucianacanteros@gmail.com

Abstract. Para realizar inferencia y generación de textos con grandes modelos de lenguaje entrenados con bases de datos que contienen sentencias judiciales y documentos legales es fundamental garantizar la confidencialidad de los datos personales y la protección de información sensible. En este trabajo, proponemos una metodología para la anonimización de bases de datos legales basada en la extracción de entidades mediante modelos de lenguaje avanzados. Se utilizaron dos modelos de lenguaje de código abierto, LLaMA 3.1 (8B) y Qwen 2.5 (7B). Cada modelo de lenguaje es entrenado en dos etapas, primero un preentrenamiento continuo en el cual se adapta el modelo al lenguaje jurídico, mejorando su capacidad de comprensión y generación de textos en este dominio especializado. Para esto se utilizó un corpus de más de 26.000 documentos legales y se evalúa la efectividad del preentrenamiento a través de métricas como BLEU, BERTScore y perplejidad. En una segunda etapa se realiza un *finetuning* específico para la tarea de anonimización y extracción de entidades. Dicho *finetuning* se evaluó en un conjunto de 50 segmentos de prueba, obteniendo un 92,79% de anonimización correcta con Qwen 2.5 (7B) y 91,58% con LLaMA 3.1 (8B), mejorando en un 4,73% y un 12,87% con respecto al modelo base con *finetuning*, respectivamente, destacando el impacto del *continued pretraining* como paso previo. Ambos entrenamientos, tanto el *continued pretraining* como el *finetuning*, fueron realizados mediante *LoRA*.

Keywords: Anonimización, extracción de entidades, continued pretraining, finetuning, dominio legal

Anonymization of Legal Documents using Large Language Models with Continued Pretraining and Finetuning

Sofia Ornella Ortman*[1,2], Luciana Belen Canteros*[1,2], Francisco Vargas[2], Gaston Escalante[2], Alejandro González Coene[2], Manuel Pulido[1,3]

[1] FaCENA, Universidad Nacional del Nordeste, Corrientes
[2] Legalhub S. A., Buenos Aires
[3] Instituto de Modelado e Innovación Tecnológica, CONICET, Argentina
ortmansofia00@gmail.com, 2010lucianacanteros@gmail.com

Abstract. To perform inference and text generation with large language models trained on datasets containing court rulings and legal documents, it is essential to ensure the confidentiality of personal data and the protection of sensitive information. In this work, we propose a methodology for the anonymization of legal databases based on entity extraction using advanced language models. Two open-source language models, LLaMA 3.1 (8B) and Qwen 2.5 (7B) are evaluated. Each language model is trained in two stages: first, a continued pretraining phase in which the model is adapted to legal language, improving its ability to understand and generate text in this specialized domain. With this end, we use a corpus of more than 26,000 legal documents composed of legislation, legal doctrine, and case law. The impact of the pretraining phase is evaluated with metrics such as BLEU, BERTScore, and perplexity. In the second stage, a task-specific finetuning is performed for anonymization and entity extraction. This finetuning is conducted using a dataset consisting of 150 segments. The finetuning was evaluated on a test set of 50 segments, achieving 92.79% correct anonymization with Qwen 2.5 (7B) and 91.58% with LLaMA 3.1 (8B), improving by 4.73% and 12.87% respectively compared to the base model with finetuning, highlighting the influence of continued pretraining as a preliminary step. Both training phases, continued pretraining and finetuning, were conducted using LoRA.

Keywords: anonymization, entity extraction, continued pretraining, finetuning, legal domain.

1 Introducción

Uno de los principales impedimentos para la adopción de grandes modelos de lenguaje (LLMs) por parte de empresas e instituciones es la presencia de bases de datos que contienen información personal sensible, la cual debe ser resguardada por cuestiones éticas y normativas legales. En particular, si se desea entrenar LLMs con estas bases de datos, es esencial implementar mecanismos de protección de información previos al entrenamiento, para garantizar que, durante la fase de inferencia, el modelo no genere afirmaciones basadas en datos personales. Esto no solo responde a buenas prácticas éticas en el uso de la inteligencia artificial, sino también al cumplimiento de marcos regulatorios como la Ley 25.326 de Protección de Datos Personales en Argentina, que exige el tratamiento confidencial de la información y el consentimiento informado de los titulares.

En este trabajo proponemos una etapa o capa previa al entrenamiento del modelo en la cual se busca exclusivamente la anonimización de la base de datos. Para esto se utilizan modelos de lenguaje que identifiquen todas las entidades ligadas a datos personales. Esta capa inicial de anonimización se puede realizar en forma offline con estrictas normas de confidencialidad y seguridad basada en encriptamientos de los datos. Una vez obtenida la base de datos anonimizada, en una segunda etapa, ésta es utilizada para el entrenamiento y la inferencia del LLM en las aplicaciones específicas de interés para la empresa o institución. De este modo, el proceso de anonimización es solo incorporado dentro del preprocesamiento de datos.

Este trabajo se enmarca en el ámbito legal, el cual es uno de los ejemplos mas paradigmáticos de la necesidad de resguardar la información personal. En general las bases de datos legales estan conformadas por sentencias judiciales, contratos, acuerdos, y demás documentos legales. En esto documentos es necesario identificar todo tipo de entidades existentes, con algunos tipos que pueden estar predefinidos, tales como nombres, direcciones, teléfonos, sin embargo de acuerdo al contexto del contrato o la sentencia pueden surgir nuevos tipos de entidades que es necesario resguardar. La anonimización de documentos legales es fundamental para utilizar grandes modelos de lenguaje en bases de datos de jurisprudencia o documentos legales para *chatbot* legales u otras aplicaciones jurídicas.

En los últimos años, se han desarrollado diversas metodologías orientadas a la extracción de entidades en textos en español, como BETO-NER en Romero et al., 2020, XLM-RoBERTa-NER en MMG, 2020 y Gliner en Zaratiana et al., 2024. Dentro de estas herramientas, Gliner se destaca por su flexibilidad para definir las entidades que se desean identificar. No obstante, su principal limitación es una ventana de contexto reducida, lo que dificulta su aplicación en textos extensos. Además, estas metodologías están orientadas únicamente a la extracción de entidades y no realizan la anonimización de los textos, lo que requiere un procesamiento adicional.

En Vargas et al., 2024, se utiliza LLaMA 2 para la extracción de entidades específicas en sentencias judiciales para el posterior procesado estadístico de interés para empresas aseguradoras. Se demuestra que el modelo base de 7B tiene

un alto porcentaje de alucinaciones para la extracción de entidades nominales en documentos legales, sin embargo a partir de un sintonizado fino se reduce substancialmente la cantidad de alucinaciones del modelo (Vargas et al., 2025). Esta metodología podría ser aplicable también para la anonimización pero en Vargas et al., 2024 no se evaluó la extracción de entidades referidas a datos personales.

A diferencia de las metodologías de anonimización existentes como Gliner (Zaratiana et al., 2024) y XLM-RoBERTa-NER(MMG, 2020) con ventanas de contexto acotadas, nuestro enfoque aprovecha LLMs de código abierto para la anonimización con ventanas de contexto mayores, como LLaMA 3.1 de 8B (Meta-AI, 2024a) y Qwen 2.5 de 7B (Qwen-Team, 2024b). Ambos modelos permiten el procesamiento eficiente de documentos largos y pueden ejecutarse en entornos *offline* de bajos recursos, lo cual es esencial para el resguardo de los datos confidenciales y especialmente relevante en aplicaciones legales. En el caso del poder judicial o juzgados, no solo es requerido el proceso de anonimización en forma *offline* para la generación de la base de datos anonimizada sino también durante el proceso de inferencia del modelo para en ambos casos resguardar la confidencialidad de la información.

Para adaptar estos modelos al dominio legal, realizamos dos etapas de entrenamiento. En primer lugar, aplicamos *continued pretraining* (Ke et al., 2023) sobre un corpus de más de 26.000 documentos jurídicos, lo que permitió adaptar el conocimiento general del modelo al vocabulario propio del lenguaje legal. Posteriormente, realizamos un *finetuning* específico para la tarea de extracción y anonimización de entidades, utilizando un conjunto de datos desarrollado con control de calidad manual.

Este enfoque se inspira en trabajos recientes que exploran la adaptación de LLMs al dominio legal mediante *continued pretraining*, como Colombo et al., 2024 con el modelo SaulLM-7B, Niklaus et al., 2025 con el modelo FLawN, y (Valerio et al., 2024). En este último, aplicaron esta metodología al modelo LLaMA 3.1 con LoRA para entrenarlo en el ámbito jurídico italiano.

A su vez, nos basamos en Chen et al., 2024 para abordar la problemática del olvido catastrófico.

En las secciones siguientes del trabajo, detallamos el preprocesamiento de la base de datos (Sección 2.1), la selección de los modelos (Sección 2.2), el procedimiento de entrenamiento (Sección 2.3) y el análisis de los resultados obtenidos (Sección 3).

2 Materiales y Métodos

2.1 Bases de Datos

La base de datos utilizada fue provista por *International Legal Group, IJG,* y se encuentra conformada por 26922 documentos legales en español de tres categorías, siendo éstas, jurisprudencias, legislaciones y doctrinas.

Continued pretraining Para la utilización de la base de datos se realizó un preprocesamiento de los datos. En primer lugar, se llevó a cabo un proceso de limpieza donde se extrajo el texto de los archivos que originalmente se encontraban en formato *HTML*, se eliminaron caracteres especiales y espacios innecesarios mediante expresiones regulares, con el objetivo de obtener texto plano.

En la Tabla 1 se presentan las estadísticas de los *tokens* de los modelos elegidos diferenciados por las distintas categorías de documentos, este análisis nos permite cuantificar el tamaño de los textos de la base de datos. Para la elaboración del conjunto de datos de entrenamiento, se descartaron todos aquellos documentos con menos de 300 *tokens* o que superaban la ventana de contexto de los modelos (128k *tokens* para LLaMA 3.1 8B y 131k tokens para Qwen 2.5 7B), finalizando con un total de 21.292 elementos. El límite inferior fue establecido para descartar documentos que no hayan sido preprocesados de manera adecuada.

Table 1. Estadísticas de *tokens* por categoría y modelo. Se presentan los valores promedio, mediana, máximo y mínimo para cada modelo (LLaMA 3.1 y Qwen 2.5) según el tipo de documento.

	Jurisprudencias		**Legislaciones**		**Doctrinas**	
	LLaMA	Qwen	LLaMA	Qwen	LLaMA	Qwen
Promedio	6328	6610	3533	3592	10577	10793
Mediana	4308	4516	1077	1113	7584	7709
Máximo	481106	513185	216726	225003	212177	221191
Mínimo	7	6	24	23	103	114

La segunda etapa de preprocesamiento, de la base de datos para el *pretraining*, consistió en la anonimización de datos personales mediante la utilización de un modelo especializado en la extracción de entidades y con el posterior reemplazo de estas entidades por etiquetas asociadas a un número, cumpliendo la función de identificadores únicos para cada entidad extraída.

Los datos anonimizados incluyeron: nombres de personas, direcciones, correos electrónicos, números de teléfono, matrículas, documentos de identidad y nombres de personas jurídicas.

Finalmente, se utilizaron 21.292 documentos legales en español anonimizados para el preentrenamiento. Como medida preventiva ante el olvido catastrófico (Winata et al., 2023), el cual es la pérdida de capacidades y conocimiento del modelo producido al continuar con su entrenamiento el cual sobrescribe su entrenamiento anterior, se emplearon dos técnicas para crear el *dataset* de entrenamiento: la combinación de nuestro propio *dataset* con el *WikiCAT dataset* (PlanTL-GOB-ES, 2022), y una reorganización total de las ubicaciones de los datos, distribuidos aleatoriamente. De esta manera, se finalizó con un total de 31.156 elementos.

Finetuning Para llevar a cabo el *finetuning para la anonimización*, se confeccionó un *dataset* propio, consistente en 200 archivos `json` con el texto sin anonimizar, el texto anonimizado y un diccionario con las entidades encontradas junto con sus correspondientes etiquetas.

Para el entrenamiento, se tomó el texto sin anonimizar como entrada al modelo. Por ejemplo,

```
‘‘En tal estado, los actores dijeron haber concurrido a otro service
    oficial Empresa Ficticia Uno S.A. ...’’
```

En cuanto a la salida esperada, se utilizó tanto el texto anonimizado como el diccionario de entidades. En la salida del modelo se espera un `json` con el siguiente formato:

```
{"texto_anonimizado": "En tal estado, los actores dijeron haber
    concurrido a otro service oficial ["company_1"] ...",
  "entities": [{"entity": "Empresa Ficticia Uno S.A.",
                "types": ["company_1"]}]}
```

Para la generación de este *dataset*, se tomaron fragmentos de textos legales correspondientes a nuestro corpus de datos y se utilizó el modelo LLaMA 3 de 70B (meta-LLaMA, 2022) para extraer las entidades que se debian anonimizar, dandoles el formato adecuado y anonimizandolas. Posteriormente, los datos pasaron por un riguroso proceso de corrección y validación manual.

Del corpus total mencionado se destinaron 150 elementos para el proceso de *finetuning*, mientras que los 50 restantes fueron asignados para *testing*.

2.2 Modelos

Para la implementación de este proyecto se adoptaron los modelos Qwen 2.5 de 7B (Qwen-Team, 2024b) y LLaMA 3.1 de 8B (Meta-AI, 2024a). La elección de estos modelos de código abierto se basó en su relevancia actual y su desempeño en distintos *leaderboards*, tanto generales como específicos del ámbito legal (Guha et al., 2023 y LMArena, 2024). En el caso de LLaMA, se consideró la experiencia y resultados existentes en el ámbito legal habiendo sido este modelo utilizado en investigaciones previas (Vargas et al., 2024 y Vargas et al., 2025). Una ventaja adicional de estos modelos es su amplia ventana de contexto: 128k *tokens* en el caso de LLaMA y 131k en el de Qwen, lo cual resulta fundamental considerando el tamaño de los documentos procesados. La selección del número de parámetros de los modelos estuvo, por un lado, condicionada por los recursos computacionales disponibles en el clúster donde se llevó a cabo el proyecto (CECONEA) pero fundamentalmente por sus posteriores aplicaciones.

2.3 Entrenamiento

El proceso de entrenamiento se dividió en dos etapas, aplicándose la misma metodología para ambos modelos. Durante la primera etapa se realizó un *continued pretraining* buscando especializar los grandes modelos prentrenados al

dominio legal. En general los LLMs entrenados con textos de diversos campos y disciplinas generales tienen un desempeño relativamente bajo en dicho dominio. Esto se explica en que el lenguaje jurídico tiene un léxico particular con numerosos latinismos, palabras a la que se les da significados propios del ámbito y fundamentaciones con contextos muy extendidos. Por estas razones, para que el modelo comprenda y genere textos legales con precisión se requiere del preentrenamiento del modelo en una base de datos específica del dominio legal. En este solo se obtiene la capacidad de predecir el próximo *token* por el modelo; sin embargo, esta permite obtener una comprensión de los contextos legales en las capas internas del modelo. Durante una segunda etapa se efectuó un *finetuning* con el objetivo de especializar ambos modelos en la detección de entidades nombradas (NER) y la anonimización del texto.

En ambas etapas se usó la librería *unsloth* (Han and team, 2023), esta librería permite cuantizar ambos modelos a 4-bits (Qwen-Team, 2024a, Meta-AI, 2024b), permitiendo ahorrar tanto tiempo de procesamiento como recursos computacionales. A su vez, para entrenar utilizamos la metodología **LoRA** (Hu et al., 2021), que permite reducir aún más los recursos necesarios para entrenar los modelos, ya que reduce los pesos en matrices representativas de los mismos, las cuales tienen menores requerimientos de memoria para ser alojadas y trabajadas.

Continued pretraining Con el objetivo de mejorar el rendimiento de los modelos se optó por una etapa de *continued pretraining* para que sean entrenados en el dominio legal. Para ello, se utilizó el 81% del *dataset* preprocesado detallado en la Sección 2.1 como subconjunto de entrenamiento, un 10% para el test final y el 9% restante para validación.

Se llevó a cabo la hiper-optimización de los siguientes hiperparámetros: *batch size*, número de épocas, *learning rate* y los correspondientes a la adaptación de bajo rango, *rank* y *alpha*, tomando como métrica la función de pérdida por entropía cruzada (*cross-entropy loss*) aplicada en los datos de validación durante el entrenamiento. Se tomaron como base los parámetros utilizados en trabajos previos (Niklaus et al., 2025; Valerio et al., 2024), y se realizó la hiperoptimización para encontrar aquellos que obtuvieron la mejor métrica. Finalmente, los parámetros seleccionados son los que se muestran en la Tabla 2.

Table 2. Hiperparámetros seleccionados para el entrenamiento de los modelos LLaMA 3.1 8B y Qwen 2.5 7B con LoRA.

Modelo	Epochs	Batch Size	Lr	LoRA Rank	LoRA Alpha
LLaMA 3.1 8B	2	4	1e−5	128	32
Qwen 2.5 7B	1	2	3e−5	16	32

El entrenamiento se realizó en el servidor del CECONEA, contando con 2x NVIDIA A40 GPUs, 48GB GDDR5 de memoria por GPU es decir un total de

96GB de memoria VRAM disponible. El tiempo de duración del entrenamiento fue de aproximadamente 20 horas para LLaMA y 14 horas para Qwen, en las cuales el uso máximo de memoria con el modelo cuantizado fue de 27 GB y 23 GB, respectivamente, representando un 28% y 24% de la memoria total disponible.

Finetuning En el caso del *finetuning* se utilizaron los mismos parámetros para ambos modelos, se entrenó con un *learning rate* de 2e-4, número de épocas 2, *batch size* de 4 y acumulación de gradientes cada 4 pasos. En cuanto a la configuración de *LoRA* se empleó un *alpha* de 32 con un *rank* de 128. El tiempo estimado de *finetuning* para ambos modelos rondó los 12 minutos, utilizando el mismo servidor mencionado para el *continued pretraining* y *Unsloth* como *framework*, obteniendo un uso máximo de memoria de 17 GB durante el proceso, representando un 18% de la memoria total disponible.

3 Evaluación y Resultados

3.1 Continued Pre-training

Métricas Para la evaluación del *continued pre-training* adoptamos el enfoque presentado en Valerio et al., 2024, el cual consiste en un análisis mediante instrucciones parciales. Los textos del conjunto de prueba se introducen de manera parcial dentro del *prompt*, los primeros 30 *tokens*, para luego comparar el resultado generado por el modelo a partir de estos primeros *tokens* con el texto original completo.

Para dicha evaluación entre el texto generado y el original se usan las siguientes métricas:

- **BLEU** (*Evaluación Bilingüe por Sustituto*) es una métrica que evalúa la similitud entre el texto generado por un modelo y el texto original, calculando la superposición de n-gramas entre ambos textos. Los valores resultantes oscilan entre 0 y 1, siendo los más cercanos a 1 los que reflejan mayor coincidencia (Papineni et al., 2002).
- **BERTScore** es una métrica de evaluación automática para la generación de texto que mide la similitud semántica entre los textos generados y los textos de referencia, utilizando representaciones contextuales obtenidas mediante el preentrenamiento de BERT (Zhang et al., 2020). Una vez obtenidos los *embeddings* de cada palabra en ambos textos, se calcula la similitud coseno entre todos los pares posibles de tokens. A partir de estas similitudes, se computan las métricas de *recall*, *precision* y *F1*. El *recall* corresponde al promedio de las máximas similitudes de cada token del texto de referencia respecto del texto generado; la *precision*, al promedio de las máximas similitudes de cada token del texto generado respecto del texto de referencia; y el *F1* se calcula como la media armónica entre ambas métricas. En términos interpretativos, el *recall* evalúa cuánto del texto de referencia está presente

en el texto generado, la *precision* evalúa cuánto del texto generado tiene respaldo en la referencia, y el *F1* refleja el equilibrio entre ambos aspectos.
- **Perplejidad** es una de las métricas más utilizadas para evaluar modelos de lenguaje. Mide la probabilidad de que una palabra sea elegida como la siguiente en la secuencia. Es decir, esta métrica cuantifica la incertidumbre del modelo. Una menor perplejidad indica un mejor rendimiento.

Table 3. Resultados de las métricas para la evaluación del impacto del *continued pretraining* en el modelo Qwen 2.5 7B y LLaMA 3.1 8B. Se muestran los valores de BLEU, BERTScore (F1, Recall y Precisión) y perplejidad. Nota: CPT: *Continued Pretraining*

Modelo	**BLEU**	**BERTScore**			**Perplejidad**
		F1	Recall	Precision	
Qwen 2.5 7B	0.0941	0.7867	0.7836	0.7905	**6.9452**
LLaMA 3.1 8B	0.1010	0.7905	0.7868	0.7948	10.7078
Qwen 2.5 7B + CPT	**0.1482**	**0.8172**	**0.8159**	**0.8188**	31.63
LLaMA 3.1 8B + CPT	0.1293	0.8124	0.8114	0.8137	31.8444
Qwen 2.5 7B + WikiCAT + CPT	0.1402	0.8139	0.8150	0.8132	21.0682
LLaMA 3.1 8B + WikiCAT + CPT	0.1181	0.8096	0.8100	0.8094	43.6162

Resultados El *continued pre-training* realizado sobre el modelo Qwen 2.5 7B y sobre el LLaMA 3.1 8b permite obtener una mejora en ambos modelos en los valores del *BLEU*, como se muestra en la Tabla 3, lo cual indica un aumento en la similitud entre el texto generado y el texto de referencia. La Tabla 3 también muestra una mejora en los valores de *BERTScore* que resultan del *continued pre-training*. Esto sugiere que también hubo un incremento en la similitud contextual entre ambos textos. Si se compara el impacto obtenido por el *continued pre-training* entre los modelos, este tiene un mayor impacto en Qwen 2.5 7B obteniéndose un *BLEU* y *BERTScore* que supera al LLaMA 3.1 8b preentrenado mientras en el caso de los modelos base, el LLaMA 3.1 8b es el que da mejores métricas. Este resultado es robusto presentándose en las cuatro medidas, *BLEU*, *F*1, *recall* y *precision*.

Por otro lado, la perplejidad tiene un aumento significativo con respecto al valor obtenido sin *continued pre-training* en ambos modelos (Tabla 3). Esto indica que el modelo presenta una mayor incertidumbre al momento de seleccionar el siguiente *token*. Las causas de este resultado no son del todo claras. El aumento de la perplejidad no se corresponde con el incremento de las métricas anteriores (*BLEU BERTScore*). Una hipótesis es que puede deberse al olvido catastrófico, sin embargo se utilizó la metodología propuesta en (Chen et al., 2024) para disminuir este problema. Para evaluar este punto, también se realizó un pretraining sin incluir a la base wikiCAT, utilizada para mitigiar el olvido catastrofico, y los resultados de perplejidad no tienen un cambio significativo,

tal como se puede observar en la Figura 1. En este sentido, en Öncel et al., 2024 se demuestra que en las tareas de adaptación de dominio, la perplejidad tiende a aumentar cuando el dataset de preentrenamiento es similiar al dataset con el cual se entrenó el modelo base.

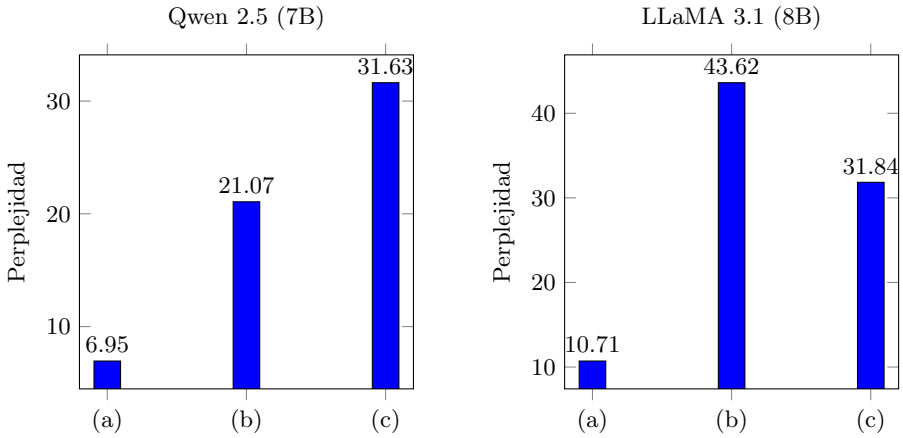

Fig. 1. Valores de perplejidad obtenidos en los diferentes experimentos realizados con los modelos Qwen 2.5 (7B) y LLaMA 3.1 (8B): (a) modelo base; (b) continued pretraining con la adición del dataset WikiCAT; y (c) continued pretraining sin el dataset WikiCAT.

3.2 Finetuning

Métricas Para evaluar el rendimiento del modelo tras el proceso de *finetuning*, se utilizó el conjunto de prueba correspondiente del *dataset*, el cual contiene segmentos previamente anonimizados y sus respectivas entidades validadas. El procedimiento consistió en introducir al modelo los textos originales sin anonimizar y comparar las entidades extraídas por el modelo con las existentes en el conjunto de validación.

Para la evaluación de los resultados del modelo, estos se clasificaron en dos grandes grupos:

- **Salidas con formato correcto**: como se mencionó anteriormente, se espera que el modelo entregue las entidades detectadas y el texto anonimizado en formato *json* . Los resultados que cumplan con esta condición se consideran con formato correcto. En algunos casos, la salida no era del todo correcta, ya que venía acompañada de un texto adicional innecesario, por ejemplo: repetir la respuesta en el formato correcto una y otra vez hasta completar el número de *tokens* para la salida. Para solucionar esto se aplicó un filtro de limpieza utilizando *regex* para extraer únicamente la respuesta *json*. Este comportamiento se observo principalmente en el modelos de LLaMA 3.1 8B.

– **Salidas con formato incorrecto**: son aquellos resultados que no se ajusten al formato esperado y no son tenidos en cuenta para el cálculo de las métricas de desempeño en anonimización.

Esta diferenciación permite calcular el porcentaje de salidas con formato correcto, que se define como:

$$\text{Porcentaje de formato correcto} = \frac{N_{\text{correcto}}}{N_{\text{correcto}} + N_{\text{incorrecto}}} \tag{1}$$

donde N_{correcto} representa la cantidad de salidas con formato válido, y $N_{\text{incorrecto}}$ la cantidad de salidas descartadas por no cumplir con el formato esperado.

A continuación, se consideran únicamente las respuestas del modelo que poseen formato válido (*json*) y se analizan las entidades detectadas, clasificándolas en los siguientes grupos:

– **Entidades totales (T)**: cantidad total de entidades presentes en el *dataset* de prueba, que deberían ser detectadas por el modelo.
– **Entidades correctas (C)**: se consideran aquellas entidades anonimizables que están presentes en el texto de referencia y que han sido correctamente identificadas por el modelo. Una entidad se considera correctamente detectada cuando existe una coincidencia exacta o cuando el *partial ratio* entre la entidad predicha y la de referencia supera un umbral de 80. Esta metodología se adoptó con el objetivo de no penalizar al modelo en casos donde identifica la entidad correctamente pero incluye información adicional, como por ejemplo *"Dr. Juan"* en lugar de simplemente *"Juan"*.
– **Entidades faltantes (F)**: entidades anonimizables presentes en el texto que no fueron detectadas por el modelo.
– **Entidades extras (E)**: elementos del texto extraidos por el modelo que no corresponden a una entidad anonimizable, pero que el modelo detectó erróneamente como tal. Ya sea que anonimizó un tipo de entidad nueva, por ejemplo: el modelo detecto como 'med' (medicina) a 'ibuprofeno', o extrajo una entidad erronea de un tipo existente, por ejemplo se detectó 'El juez' como una entidad de tipo 'persona'. Notar que en el primer caso considerando el modelo genera un nuevo tipo de entidad se puede filtrar facilmente, sin embargo en el segundo caso no se pueden identificar tan sencillamente.

Esta clasificación permite definir las siguientes métricas para el análisis del desempeño del modelo:

– **Porcentaje de anonimización**: representa la proporción de entidades correctamente identificadas por el modelo (C) con respecto al total de entidades presentes en el texto de referencia (T). Se calcula como:

$$\text{Porcentaje de anonimización} = \frac{C}{T} \tag{2}$$

– **Porcentaje de entidades extras**: mide la proporción de entidades detectadas por el modelo que no están presentes en el texto de referencia (E), en relación con el total de entidades esperadas (T). Se define como:

$$\text{Porcentaje de entidades extras} = \frac{E}{T} \tag{3}$$

Table 4. Resultados de evaluación para la tarea de extracción de entidades y anonimización. Se presentan los resultados para dos modelos base, con y sin *continued pretraining*, seguidos de *finetuning*. Nota: FT: *Finetuning*, CPT: *Continued Pretraining*

Modelo	**% Formato Correcto**	**% Anonimización**	**% Extras**
LLaMA 3.1 8B + FT	**100%**	78.71%	11.25%
LLaMA 3.1 8B + CPT + FT	88.33%	91.58%	4.52%
Qwen 2.5 7B + FT	88.33 %	88.06%	3.80%
Qwen 2.5 7B + CPT + FT	92.5%	**92.79%**	**1.45%**

Resultados En la Tabla 4 se presentan los resultados de la evaluación de la anonimización resultante del *finetuning* aplicado a los modelos Qwen 2.5 7B y LLaMA 3.1 8B, tanto en su versión base como con el modelo que fue preentrenado previo al *finetuning*. Para esta evaluación se utilizó un total de 120 entidades pertenecientes al *dataset* de *testing* de 50 fragmentos (2.1), en primera instancia se evaluó si la respuesta del modelo respetaba el formato esperado (*json*) y luego, en la segunda, se midieron los porcentajes de anonimización y extras. La cantidad de entidades a anonimizar varía por cada experimento debido a que no todas las respuestas del modelo pasaron la primera instancia, sin embargo, esto no influye en ninguna medida para los porcentajes de anonimización y extras obtenidos en la segunda instancia, es decir, aquellas entidades que corresponden a respuestas que no obedecen el formato no clasifican como entidades faltantes F.

En cuanto al porcentaje de formato correcto, únicamente el modelo de Llama base ($LlaMa3.18B + FN$) lo respetó en su totalidad, resultado que se vio afectado luego del entrenamiento con CPT.

La implementación de un *continued pretraining* previo al *finetuning* tiene un importante impacto en las métricas (Tabla 4). Este mejora el porcentaje de anonimización en un 4.73% en el caso de Qwen 2.5 7B y en un 12.87% para LLaMA 3.1 8B. Así como disminuye el porcentaje de entidades extras en un 2.35% y un 6.73%, respectivamente. Este último resultado es relevante, ya que si el porcentaje de entidades extras es alto, estaríamos etiquetando muchas partes del texto como entidades que no lo son, afectando la coherencia del mismo.

Finalmente, podemos concluir que el modelo con mejor desempeño es Qwen 2.5 7B con *continued pretraining* más *finetuning* para la tarea de extracción de entidades y anonimización de textos legales, alcanzando un porcentaje de anonimización del 92.79%. Por otro lado es el que tiene el menor porcentaje de extras, 1,45%.

4 Conclusión

El dominio legal posee un lenguaje con características específicas, con vocabulario técnico, latinismos y fundamentaciones con contextos muy extendidos. En este trabajo se realiza la adaptación al dominio legal de los modelos Qwen 2.5 7B y LLaMA 3.1 8B mediante un *continued pretraining* en una base de sentencias y documentos legales de Argentina. En una segunda etapa se realiza un *finetuning* de estos modelos para la tarea de extracción de entidades y anonimización de textos legales.

Los resultados obtenidos demuestran la utilidad del *continued pretraining* para la adaptación al dominio legal, reflejada en un aumento de las métricas de *recall*, precisión y F1 Score según BERTScore, así como en una mejora en el BLEU: de 0,0941 a 0,1402 en el caso de Qwen 2.5 (7B), y de 0,1010 a 0,1181 en el caso de LLaMA 3.1 (8B), siendo el incremento más notable en el primer modelo. Asimismo, los modelos con *continued pretraining* presentaron una mejora en el porcentaje de anonimización, con un incremento del 4,73% en Qwen 2.5 (7B) y del 12,87% en LLaMA 3.1 (8B).

El *finetuning* también mostró resultados satisfactorios, alcanzando un porcentaje de anonimización del 91,58% en LLaMA 3.1 (8B) y del 92,79% en Qwen 2.5 (7B), ambos con *continued pretraining*. Además, se observó una reducción en el porcentaje de entidades extras detectadas: de 11,25% a 4,52% en LLaMA y de 3,80% a 1,45% en Qwen. Estos resultados indican un mejor desempeño general del modelo Qwen en ambas métricas. Cabe destacar que estos experimentos se realizaron con un conjunto de datos reducido de 150 textos, por lo que consideramos que estos resultados motivan la realización de experimentos futuros con bases de mayor tamaño para evaluar su impacto en la *performance*.

Este trabajo representa un puntapié inicial para futuros desarrollos orientados al entrenamiento de modelos con bases de datos de documentos legales anonimizados, con el objetivo de facilitar el procesamiento, la generación de textos y la asistencia en tareas propias del ámbito judicial.

5 Agradecimientos

Agradecemos a *IJ International Legal Group*[4] por proveer los datos necesarios para el desarrollo de este trabajo.

También agradecemos a *LegalHub*[5] y a la Universidad Nacional del Nordeste, por brindarnos el espacio para aprender y seguir creciendo.

Este trabajo utilizó recursos computacionales provistos por el Centro de Cómputos de Alto Desempeño del Nordeste Argentino (CECONEA)[6].

[4] https://ij-ilg.com/
[5] https://legalhub.la/
[6] http://cad.unne.edu.ar/index.php

References

Chen, J., Chen, Z., Wang, J., Zhou, K., Zhu, Y., Jiang, J., Min, Y., Zhao, W. X., Dou, Z., Mao, J., Lin, Y., Song, R., Xu, J., Chen, X., Yan, R., Wei, Z., Hu, D., Huang, W., & Wen, J.-R. (2024). Towards effective and efficient continual pre-training of large language models. https://arxiv.org/abs/2407.18743

Colombo, P., Pires, T. P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A. F. T., Esposito, F., Raposo, V. L., Morgado, S., & Desa, M. (2024). Saullm-7b: A pioneering large language model for law [Accessed: 2025-04-13]. *arXiv preprint arXiv:2403.03883*. https://arxiv.org/abs/2403.03883

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., ... Li, Z. (2023). Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*. https://doi.org/10.48550/arXiv.2308.11462

Han, D., & team, U. (2023). *Unsloth*. http://github.com/unslothai/unsloth

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. https://arxiv.org/abs/2106.09685

Ke, Z., Shao, Y., Lin, H., Konishi, T., Kim, G., & Liu, B. (2023). Continual pre-training of language models [Published at ICLR 2023]. *arXiv preprint arXiv:2302.03241*. h

LMArena. (2024). Chatbot arena llm leaderboard: Community-driven evaluation for best llm and ai chatbots. Retrieved April 13, 2025, from https://lmarena.ai/

Meta-AI. (2024a). Llama-3.1-8b. Retrieved April 13, 2025, from https://huggingface.co/meta-LLaMA/LLaMA-3.1-8B

Meta-AI. (2024b). Meta-llama-3.1-8b-bnb-4bit. Retrieved April 13, 2025, from https://huggingface.co/unsloth/Meta-LLaMA-3.1-8B-bnb-4bit

meta-LLaMA. (2022). Meta-llama-3-70b. Retrieved April 13, 2025, from https://huggingface.co/meta-LLaMA/Meta-LLaMA-3-70B

MMG. (2020). Xlm-roberta-large-ner-spanish. Retrieved April 13, 2025, from https://huggingface.co/MMG/xlm-roberta-large-ner-spanish

Niklaus, J., Zheng, L., McCarthy, A. D., Hahn, C., Rosen, B. M., Henderson, P., Ho, D. E., Honke, G., Liang, P., & Manning, C. (2025). Lawinstruct: A resource for studying language model adaptation to the legal domain. *arXiv preprint arXiv:2404.02127*. https://arxiv.org/abs/2404.02127

Öncel, F., Bethge, M., Ermis, B., Ravanelli, M., Subakan, C., & Yıldız, Ç. (2024). Adaptation odyssey in llms: Why does additional pretraining sometimes fail to improve? *arXiv preprint arXiv:2410.05581*. https://arxiv.org/abs/2410.05581

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation [IBM T.J. Watson Research Center]. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318. https://doi.org/10.3115/1073083.1073135

PlanTL-GOB-ES. (2022). Wikicat$_e$*sv*2. Retrieved April 13, 2025, from https://huggingface.co/datasets/PlanTL-GOB-ES/WikiCAT_esv2

Qwen-Team. (2024a). Qwen2.5-7b. https://huggingface.co/unsloth/Qwen2.5-7B

Qwen-Team. (2024b, September). Qwen2.5: A party of foundation models. https://qwenlm.github.io/blog/qwen2.5/

Romero, M., Tomeh, N., Holat, P., & Charnois, T. (2020). Bert-spanish-cased-finetuned-ner. Retrieved April 13, 2025, from https://huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner

Valerio, F., Basile, P., & de Gemmis, M. (2024). Adapting a large language model to the legal domain: A case study in italian [Department of Computer Science, University of Bari Aldo Moro; AI2B srl, Spin-Off of the University of Bari Aldo Moro]. *arXiv preprint arXiv:2403.20007*. https://arxiv.org/abs/2403.20007

Vargas, F., Gonzalez Coene, A., Escalante, G., Lobón, E., & Pulido, M. (2024). Extracción de entidades en sentencias judiciales usando llama-2. *ASAID - Simposio Argentino de Inteligencia Artificial y Ciencia de Datos, JAIIO*. https://publicaciones.sadio.org.ar/index.php/asaid/article/view/2999

Vargas, F., González Coene, A., Escalante, G., Lobón, E., & Pulido, M. (2025). El impacto del ajuste fino de llama en las alucinaciones para la extracción de entidades nominales en documentos legales. *SADIO Electronic Journal of Informatics and Operations Research*, *24*(1). https://doi.org/10.24215/15146774e068

Winata, G., Xie, L., Radhakrishnan, K., Wu, S., Jin, X., Cheng, P., Kulkarni, M., & Preotiuc-Pietro, D. (2023, July). Overcoming catastrophic forgetting in massively multilingual continual learning. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 768–777). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.48

Zaratiana, U., Tomeh, N., Holat, P., & Charnois, T. (2024). Gliner: Generalist model for named entity recognition using bidirectional transformer. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5364–5376. https://doi.org/10.18653/v1/2024.naacl-long.300

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert [To appear in ICLR 2020]. *arXiv preprint arXiv:1904.09675*. https://doi.org/10.48550/arXiv.1904.09675