# Summarization of Legal Documents using Large Language Models with Fine-tuning

Luciana Belen Canteros[1], Smulever Federico[1,2], Francisco Vargas[1], Sonia I. Mariño[1], and Manuel Pulido[1,2]

[1] FaCENA, Universidad Nacional del Nordeste, Corrientes
[2] Instituto de Modelado e Innovación Tecnológica, CONICET, Argentina
luciana.canteros@comunidad.unne.edu.ar

**Abstract.** Summarizing extensive legal documents is crucial for optimizing professional workflows in the legal domain and enhancing the precision of Retrieval-Augmented Generation (RAG) systems. This work proposes a methodology based on fine-tuning open-source Large Language Models (LLMs), specifically Llama and Qwen, integrated with various prompting techniques. The models are trained and evaluated on a substantial corpus comprising over fourteen thousand Argentine judicial rulings. The objective is to develop models specialized in abstractive summarization while maintaining strict control over output token length. The results demonstrate that fine-tuned models, combined with optimized prompting techniques, capture relevant content more effectively and maintain higher semantic coherence than base models. Notably, the Llama 3.1 Instruct (8B) model achieved a relative improvement in ROUGE L of 63.91% over its base version, while the same model achieved a relative improvement in BERTScore F1 of 12.66%. The resulting technique will be integrated into RAG systems and used to generate legal databases where high-quality summaries serve as essential metadata.

**Keywords:** Fine-tuning, abstractive summary, LLMs, legal domain

# Generación Automática de Resúmenes de Documentos Legales utilizando Grandes Modelos de Lenguaje con Fine-tuning

Luciana Belen Canteros[1], Smulever Federico[1,2], Francisco Vargas[1], Sonia I. Mariño[1], and Manuel Pulido[1,2]

[1] FaCENA, Universidad Nacional del Nordeste, Corrientes
[2] Instituto de Modelado e Innovación Tecnológica, CONICET, Argentina
luciana.canteros@comunidad.unne.edu.ar

**Abstract.** La confección de resúmenes a partir de documentos extensos del ámbito legal son de importancia para optimizar la tarea de los profesionales del dominio, como también para contribuir a una extracción de mayor calidad y precisión en sistemas *RAG*. En este trabajo se propone una metodología basada en un entrenamiento de ajuste fino (*fine-tuning*) de los modelos de lenguaje de código abierto, Llama y Qwen, combinada con diversas técnicas de prompting. El entrenamiento y la evaluación son aplicados en un corpus de más de catorce mil sentencias de la jurisprudencia argentina. El objetivo es obtener modelos especializados en resúmenes abstractivos, mientras se prioriza mantener una generación donde la longitud de *tokens* este controlada. Los resultados obtenidos evidencian que los modelos que han pasado por un proceso de *fine-tuning* junto con el uso de diferentes técnicas de ingeniería de prompts logran capturar de manera más efectiva el contenido relevante de los documentos, manteniendo coherencia semántica y superando a modelos base mientras se respetan restricciones de longitud establecidas. En particular, el modelo Llama 3.1 Instruct (8B) alcanza mejoras relativas de ROUGE L del 63.91% respecto a su versión base y mejoras relativas del 12.66% para BERTScore F1. La técnica desarrollada será utilizada para la generación de bases de datos legales en la cual los resúmenes son parte de la metadata como también para su integración en un sistema *RAG*.

**Keywords:** fine-tuning, resumen abstractivo, grandes modelos de lenguaje, dominio legal

## 1 Introducción

Uno de los principales desafíos a la hora de trabajar en el dominio legal yace en la necesidad de manipular y analizar grandes volúmenes de textos donde, generalmente, los documentos contienen terminología especializada y expresiones complejas (Ragazzi et al., 2024).

Este proceso laborioso consume una cantidad de tiempo considerable a profesionales como abogados y jueces, tiempo que podría destinarse a otras tareas. Herramientas de *summarization* automatizadas contribuyen a reducir la dimensión de documentos legales, disminuyendo tanto el tiempo de ejecución como el costo (Akter et al., 2025). Asímismo, los resúmenes son utilizados como parte de metadata para sistemas de busqueda general de documentos y también, en la actualidad, se integran formando parte de sistemas RAG [3] (de Martim, 2025).

La generación de resúmenes busca condensar textos extensos en versiones concisas, capturando la esencia del contenido original en lugar de simplemente extraer fragmentos del mismo (Akter et al., 2025). El enfoque abstractivo emula la capacidad de síntesis humana al interpretar el documento de origen y producir un resumen basado en sus conceptos clave (Rush et al., 2015).

En la elaboración de un resumen, se considera que el nivel de detalle es insuficiente cuando la brevedad compromete la comprensión del contenido esencial. No obstante, la inclusión excesiva de datos suele dificultar la legibilidad sin incrementar proporcionalmente el valor informativo. Existe un umbral crítico en la densidad de información: superado este límite, el resumen puede volverse ininteligible o incluso incurrir en alucinaciones o errores factuales (Adams et al., 2023). Por último, un resumen cuya extensión no represente una reducción significativa respecto al documento original pierde su propósito funcional y su justificación como herramienta de síntesis.

Este proyecto fue financiado y realizado en el marco del programa de becas IA-UNNE (Eje Legal) de la Universidad Nacional del Nordeste, con el objetivo de desarrollar técnicas que serán utilizadas en la generación de bases de datos legales, en las cuales los resúmenes forman parte de la metadata y se integrarán en un sistema RAG, en conjunto con la metodología propuesta en Smulever et al., 2026.

En este trabajo presentamos una investigación comparativa sobre el ajuste de diferentes prompts, modelos y parámetros mientras se busca optimizar *LLMs* [4] dedicados a la *summarization* abstractiva de documentos legales mediante un entrenamiento con sintonizado fino, *fine-tuning*[5], utilizando una base de datos propia que contiene resúmenes desarrollados por profesionales, de forma similar a Heddaya et al., 2024.

[3] RAG, del inglés *Retrieval-Augmented Generation*, se refiere a sistemas de generación aumentada por recuperación que combinan modelos generativos con mecanismos de búsqueda de información externa.

[4] LLMs, del inglés *Large Language Models*, refiere a grandes modelos de lenguaje entrenados con grandes volúmenes de texto para procesar y generar lenguaje natural.

[5] El *fine-tuning* consiste en adaptar un modelo previamente entrenado a una tarea o dominio específico mediante entrenamiento con datos especializados.

Se ha comprobado que la adaptación de grandes modelos de lenguaje al dominio legal evidencia un impacto positivo en tareas legales específicas. En Ortman, Canteros, et al., 2025 se utiliza LLaMA 3.1 y Qwen 2.5 para la tarea de anonimización de documentos legales, constatandose como a partir del entrenamiento de modelos con *continued pretraining* y *fine-tuning* se logra una mejora en la taza de anonimización. Otros investigadores han explorado como el entrenamiento en modelos de *summarization* abstractiva integrando entidades legales ofrece una mejora en la calidad de los resúmenes, optimizando la inclusión de términología específica del medio, como Ajay Mukund and Easwarakumar, 2025 y Steffes et al., 2025.

Para la evalución de la calidad de los resúmenes generados sobre los documentos legales originales se utilizan métricas basadas en la superposición léxica como ROUGE, donde se evalua la calidad de textos generados automáticamente, comúnmente utilizada en tareas de summarization. Sin embargo, en Steffes et al., 2023 se detalla que estas métricas tradicionales no son suficientes para validar el desarrollo de resúmenes especializados en el área legal, y para prescindir de validación humana, se utilizaron métricas basadas en embeddings, en particular BERTScore, que mide similitud semántica mediante comparación de embedding contextuales generados por el modelo BERT (Akter et al., 2025).

En la sección 2.1 del trabajo, explicamos la composición de la base de datos de documentos legales y los modelos utilizados, Sección 2.2. Detalles tanto sobre el entorno de entrenamiento como de los parámetros utilizados son detallados en Sección 2.3. En Sección 3, se realiza un análisis sobre los resultados obtenidos. Mientras en la Sección 4 se presentan las conclusiones del trabajo y las posibles líneas futuras de investigación.

## 2 Materiales y Métodos

### 2.1 Base de Datos

Para el presente trabajo se uso la base de datos provista por *IJ - International Legal Group*, conformada por 26998 documentos legales en idioma español divididos en tres categorías: jurisprudencias, legislaciones y doctrinas. En la mayoria de los registros, los documentos se encuentran acompañados con un resumen desarrollado en forma manual por profesionales del área legal. Como paso previo a la confección de la base de datos para el *fine-tuning* (Sección 2.3), se realizó un preprocesamiento de los datos para obtener el texto plano de cada documento que originalmente estaba en formato *HTML*, para esto se utilizó la librería Beautiful Soup(Richardson, 2007).

En este trabajo nos enfocamos en utilizar únicamente jurisprudencias tanto para el entrenamiento (Sección 2.3) como para la validación de los modelos entrenados (Sección 3). Este tipo de documentos poseen resúmenes, antes mencionados, para el 99,80% de las sentencias, a diferencia de las doctrinas (14,57% de los casos) y las lesgislaciones (13,60% de los casos). Además, los resúmenes hallados en doctrinas y legislaciones carecian de consistencia en cuanto a su estructura,

por ejemplo, algunos se conformaban con un listado de las leyes o palabras clave nombradas durante el texto. Por otro lado, los resúmenes correspondientes a jurisprudencias eran uniformes y mantenian la misma estructura de condensación del documento original que se esperaba replicar en este proyecto.

La base posee un total de 14836 jurisprudencias de las cuales posteriormente se realizó una limpieza donde se eliminaron documentos menores a 200 palabras, documentos con resúmenes donde el texto no se encuentra completo, documentos con resúmenes del mismo tamaño o más largos que el documento original y textos inválidos en general. A partir del conjunto de jurisprudencias preprocesado y filtrado se desarrolló una base de datos en formato *CSV* con los pares jurisprudencia original y resúmen generado por experto, contando con 14612 elementos.

Para el 79.01% de los documentos que contaban con resúmenes la longitud de los resúmenes equiparaba a un 10.66% del documento original, siendo el promedio general de 263 *tokens* por resúmen, una desviación estandar del 10.49% y diferencia promedio de 70 *tokens*. Además, la mayoría de resúmenes contenían menos de 5 oraciones en las cuales la media de *tokens* resultó en 50 con una desviación 27.1.

Se optó por no filtrar documentos clasificados como demasiado extensos (aquellos que superan los 15000 *tokens*) ya que representaban un porcentaje mínimo al total, siendo este el 2.37%, es decir 348 documentos. La estrategia utilizada para documentos que excede el límite de *tokens* se encuentra detallada en Sección 2.3.

### 2.2 Modelos

Los modelos utilizados en este estudio son Qwen y LlaMA. Esta elección esta basada en que se priorizaron opciones de tipo *open source* que permitan una amplia ventana de contexto (superior a 128k *tokens*) y que se ajuste a la longitud de los textos utilizados (Sección 2.1). Además, se tuvo en cuenta para la selección los antecedentes del grupo de trabajo en estos modelos (Ortman, Canteros, et al., 2025, Vargas, Coene, et al., 2025, Vargas et al., 2024 y Vargas, González Coene, et al., 2025).

Finalmente, se optó por las versiones mas actuales de estos modelos *Llama 3.1 8B Instruct* (Meta-AI, 2024b) y *Qwen 2.5 7b Instruct* (Qwen-Team, 2024b) junto con los modelos base *Llama 3.1 8B* (Meta-AI, 2024a) y *Qwen 3 8B* (Team, 2025).

### 2.3 Entrenamiento

Para el entrenamiento de los modelos se usó la librería *unsloth* (Han and team, 2023) que permite la cuantización de los parámetros del modelo (Qwen-Team, 2024a, Meta-AI, 2024c) junto con la metodología LoRA (Hu et al., 2021), reduciendo de esta manera los recursos computacionales necesarios y tiempo requerido en el entrenamiento.

Se realizó un entrenamiento de tipo *fine-tuning* a los modelos, donde en una etapa inicial se utilizó una base de datos pequeña de 500 pares de documento/resumen mientras que en la segunda etapa se utilizó un total de 14612 elementos (Sección 2.1) reservandose 150 para la evaluación.

Se utilizaron los mismos parámetros en todos los entrenamientos, siendo estos un *learning rate* de 2e-4, número de épocas 2, *batch size* de 4 y acumulación de gradientes cada 4 pasos. Se empleó un *alpha* de 32 con un *rank* de 128 para la configuración *LoRA*. Además, para manejar los textos que excedian el máximo de *tokens*[6] se optó por truncar como se realizó en Steffes et al., 2025, por este motivo se mantuvo la opción de truncate en el *pipeline* como *True*, tanto para entrenamiento como generación.

En cuanto a los recursos computacionales para el *fine-tuning*, estos entrenamientos se realizaron en *2x NVIDIA A40 GPUs*, *48GB GDDR5* de memoria por *GPU*, con un total de 96GB de memoria *VRAM* disponibles para el modelo. El tiempo de duración para generaciones de 150 resúmenes con promedios de 300 *tokens* fue en promedio de 5 horas y 6 horas.

**Table 1.** Modelos seleccionados para entrenamiento *fine-tuning* y las respectivas bases de datos utilizadas

| Modelo | Base de Datos |
|---|---|
| Llama 3.1 8B base | 14612 |
| Qwen 3 8B base | 14612 |
| Llama 3.1 8B Instruct | 14612 |
| Qwen 2.5 7B Instruct | 14612 |

## 3 Evaluación y Resultados

Para evaluar el desempeño de los modelos pre y post *fine-tuning* se realizó una generación de resúmenes abstractivos a partir de la base de datos original de documentos legales utilizando una muestra fija de 150 documentos separados de la base de datos original. Esta generación abstractiva se realizó a partir de diferentes estrategias de *prompting*[7] propuestas y ajustes en el parámetro temperatura.

Se utilizaron tres variaciones de *prompts*, en formato *markdown*, para la generación de resúmenes abstractivos:

***Prompt* Base (PB):** *Prompt* especializado en nuestro caso específico. Se busca dar órdenes específicas y mantener una estructura definida en el *prompt*

[6] Los *tokens* son unidades mínimas de texto procesadas por el modelo, que pueden corresponder a palabras, partes de palabras, signos de puntuación u otros fragmentos según el caso.

[7] El *prompting* consiste en diseñar instrucciones de entrada para guiar el comportamiento de un modelo de lenguaje para que realice una tarea específica.

que facilite la lectura dividiendolo en tres partes. La longitud de este *prompt* es de 316 *tokens*.

***Chain of Density* original (COD):** Este técnica de prompting permite la *Chain of Density* debido a que la densificación en los resúmenes es preferida por los profesionales legales y también brinda un control de tokens durante la generación de respuestas. Adams et al., 2023. La longitud de este *prompt* es de 567 *tokens*. Se utilizó el *prompt* proveniente del trabajo original, traducido al español y con mínimas variaciones de palabras para que se ajuste al dominio legal.

***Chain of Density* re-estructurado (COD-RE):** En este caso el *prompt* original de *Chain of Density* (Adams et al., 2023) es reformulado y dividido en tres partes con el objetivo principal de garantizar un correcto formato de salida *JSON*. Para la reformulación y reescritura de cada sección, se tomaron en cuenta los trabajos Wei et al., 2023, Wang et al., 2023 y Zhou et al., 2023. La longitud de cada *prompt* es de 84, 582 y 318 *tokens* respectivamente.

Por otro lado, se buscó que los resúmenes generados mantengan un número ligeramente mayor al del promedio de *tokens* originales de los resúmenes elaborados por profesionales del dominio legal (promedio general de 263 *tokens* por resúmen de jurisprudencia en Sección 2.1), por lo que se decidió rondar los 350 tokens. Para lograr este resultado se realizaron ajustes en el *prompt* de *Chain of Density* (tanto original como re-estructurado) donde se solicitó explícitamente que contenga 4–7 oraciones con alrededor de 50 palabras al igual que los resúmenes originales.

Se calificaron como resúmenes inválidos todas aquellas respuestas que no cumplieran el formato buscado de tipo *JSON* (solo se encuentran casos de este tipo para evaluación de *Chain of Density* original) o vacías. Se mantuvo conteo de aquellos resúmenes generados con un formato incorrecto o vacíos.

Para minimizar la taza de errores de inválidos por formato incorrecto se aplicaron técnicas básicas de limpieza y recuperación, como normalización de comillas, mayúsculas, títulos en valores *JSON* que los modelos podrian reemplazar en lugar del que se les fue solicitado, etc. En el *prompt* de *Chain of Density* se solicita al modelo ejecutar 5 iteraciones, se observó que la respuesta del modelo tendía a realizar más iteraciones de las que fueron requeridas o, más crítico aún, romper el formato *JSON* a medida que la generación crecía.

Inicialmente se encontró en los experimentos de *chain of density* errores de formato y una predisposición del modelo a deteriorar las respuestas a medida que se avanzaba en las iteraciones. Esto motivó la metodología del chain of density re-estructurado en la cual se reemplazó el *prompt* de *chain of density* por tres prompts individuales, siendo estos encargados de las tareas de resúmen inicial, extracción de entidades informativas y densificación del resúmen como se visualiza en la figura 1. De esta manera, no se requiere al modelo contestar en un formato de salida específico, a excepción del *prompt* de extracción de entidades informativas donde se busca un formato de tipo lista, separando cada entidad por punto y coma. Adicionalmente, para evitar errores en la salida de la lista, se agregaron tres oportunidades de reintento en caso de que el prompt de extracción

de entidades informativas falle. Finalmente, gracias a este ajuste, se garantiza la respuesta del modelo en formato *JSON* y el control sobre la longitud de *tokens* mediante la restricción en el parámetro de generación del modelo *max tokens* por cada respuesta.


**Fig. 1.** Flujo de *Chain of Density Re-Estructurado.*

Para analizar la calidad de los resumenes generados se utilizan métricas tradicionales y basadas en embeddings de evaluación automatizada. Entre las métricas de evaluación tradicionales se usa ROUGE (Lin, 2004) y BLEU (Papineni et al., 2002). A pesar de que ROUGE es la métrica de evaluación de resúmenes más utilizada en *summarization* legal, en Steffes et al., 2023 se concluye que su uso no es lo suficientemente fiable como única métrica debido a que no mide el contenido legal de forma precisa ni exhaustiva al centrarse en la coocurrencia y la superposición de palabras. Por ello, en este trabajo se usa también una metrica basada en embeddings, BERTScore (Zhang* et al., 2020), siguiendo el enfoque adoptado en Steffes et al., 2025, a diferencia de ROUGE palabras semanticamente similares o sinónimos pueden ser detectados gracias a su cálculo de similitud de *tokens* utilizando *embeddings* contextuales de un *LLM*.

Además, se mide el impacto del parámetro del número máximo de tokens en la generación de resúmenes sobre las métricas tradicionales y BERTScore con el modelo Llama Instruct 3.1 B ajustado mediante *fine-tuning* (base de datos pares documento-resúmen original).

Se analizan los resultados obtenidos pre y post *fine-tuning* de los modelos entrenados con una generación de 150 resúmenes abtractivos a partir de documentos legales provenientes de la base de datos original para cada prueba, detallandose a continuación diferentes experimentos para: parámetro temperatura, utilización de diferentes estrategias de *prompting*, entrenamiento *fine-tuning* y parámetro *max tokens* [8] durante la generación.

En primera medida se realizaron experimentos para evaluar el impacto de la temperatura en la calidad de los resúmenes mediante las métricas BERTScore y ROUGE L con la técnica de *prompting* COD-RE. Como se puede ver en la Tabla 2 para la métrica ROUGE L no existe una correlación entre estos dos parámetros para el modelo Qwen 2.5 7B Instruct, no obstante para el modelo

[8] El parámetro *max tokens* establece el número máximo de *tokens* que el modelo puede generar en una respuesta.

Llama 3.1 8B Instruct si existe una relación entre el crecimiento de la temperatura con el aumento de la métrica ROUGE L y BERTScore, además en otros experimentos se comprobó que la generación en formato JSON para este modelo también empeoró significativamente. Sin embargo, para este último experimento antes nombrado, la caida de este valor podría estar relacionado con la posible predisposición del modelo Llama Instruct a salirse del formato con mayor facilidad a mayor temperatura. En experimentos preliminares también se evaluó el impacto de la temperatura utilizando la técnica de *prompting* COD-RE y *fine-tuning*, no se encontraron dependencias destacables en ROUGE-L y tampoco para la métrica BERTScore a pesar de las variaciones en temperatura, en todas las instancias los valores se mantuvieron estables, siendo entonces el prompting COD-RE mas robusto a la temperatura. A partir de estos resultado se concluye como temperatura óptima al valor 0.9. Para los siguientes experimentos, se decidio utilizar la temperatura 0.1 para priorizar estabilidad en las pruebas.

**Table 2.** Resultados de los experimentos para la evaluación del impacto del parámetro temperatura con COD-RE prompting. Se muestran los valores de las métricas BERTScore (Precision, Recall y F1) y ROUGE L. Notación: T: Temperatura. L: modelo Llama, Q modelo Qwen, INS: modelo Instruct.

| **Modelo** | **T** | **BERTScore** | | | **ROUGE** |
|---|---|---|---|---|---|
| | | **Precision** | **Recall** | **F1** | **ROUGE L** |
| L 3.1 8B INS | 0.1 | 0.6000 | 0.5766 | 0.5875 | 0.0859 |
| L 3.1 8B INS | 0.6 | 0.6163 | 0.5972 | 0.6061 | 0.0883 |
| L 3.1 8B INS | 0.9 | **0.6508** | **0.6306** | **0.6400** | **0.0980** |
| Q 2.5 7B INS | 0.1 | 0.6141 | 0.5925 | 0.6026 | 0.0851 |
| Q 2.5 7B INS | 0.6 | 0.6325 | 0.6112 | 0.6211 | 0.0877 |
| Q 2.5 7B INS | 0.9 | 0.6324 | 0.6111 | 0.6210 | 0.0876 |

**Table 3.** Resultados de los experimentos para la evaluación del impacto de las diferentes técnicas de *prompting* en el modelo Llama 3.1 8B Instruct . Se muestran los valores para BERTScore (Precision, Recall y F1) y ROUGE L. Notación: INV: Inválidos. PB: Prompt B, COD: Chain of Density. COD-RE: Chain of Density re-estructurado.

| **Modelo** | **INV** | **BERTScore** | | | **ROUGE** |
|---|---|---|---|---|---|
| | | **Precision** | **Recall** | **F1** | **ROUGE L** |
| PB | 0% | **0.6365** | 0.6065 | **0.6209** | 0.0780 |
| COD | 40% | 0.6238 | **0.6227** | **0.6227** | **0.1164** |
| COD-RE | 0% | 0.6000 | 0.5766 | 0.5875 | 0.0859 |

El segundo experimento se concentra en medir la sensibilidad de las tres técnicas de prompting: PB, COD, COD-RE. En la tabla 3 se encuentran los resultados

de las métricas BERTScore (Precision, Recall y F1) y ROUGE L, también se incluye una medida porcentual de los resúmenes con formatos inválidos (INV). En general la técnica de *prompting* COD es la que obtiene las mejores métricas, en particular el modelo Llama 3.1 8B Instruct obtuvo la mejor performance para ambos BERTScore y ROUGE L sin embargo este no posee control sobre el número de tokens máximo del resumen. A pesar de la mitigación de errores de formateo implementadas en el *prompting* COD, el porcentaje de respuestas que no cumplen el formato solicitado sigue siendo elevado. Además, en su mayoría estos experimentos presentaron errores en el número de iteraciones solicitadas en el *prompt* así como dificultades para mejorar la densificación. Se observa que el método de *prompting* propuesto, COD-RE, logra eliminar el número de inválidos por completo mientras provee control sobre el número de *tokens* del resúmen generado.

**Table 4.** Resultados de las métricas BERTScore y ROUGE L para la evaluación del impacto del *fine-tuning* en los modelos estudiados utilizando el método de *prompting* COD-RE. Nota: L: modelo Llama, Q: modelo Qwen, INS: modelo Instruct.

| **Modelo** | **BERTScore** | | | **ROUGE** |
|---|---|---|---|---|
| | **Precision** | **Recall** | **F1** | **ROUGE L** |
| **Resumen original** | **0.7092** | **0.6653** | **0.6860** | **0.153** |
| L 3.1 8B INS | 0.6000 | 0.5766 | 0.5875 | 0.0859 |
| Q 2.5 7B INS | 0.6141 | 0.5925 | 0.6026 | 0.0851 |
| ***fine-tuning (14612)*** | | | | |
| L 3.1 8B INS | 0.6780 | 0.6471 | 0.6619 | **0.1408** |
| Q 2.5 7B INS | **0.6938** | **0.6536** | **0.6727** | 0.1200 |
| L 3.1 8B | 0.5906 | 0.5652 | 0.5774 | 0.1152 |
| Q 3 8B | 0.6728 | 0.6378 | 0.6546 | 0.0725 |

En un tercer conjunto de experimentos se evalua el impacto del *fine-tuning* en la calidad de los resúmenes a través de las métricas BERTScore y ROUGE L en la tabla 4, ajustando los modelos descriptos con una base de datos de 14612 elementos y realizando una generación de resúmenes utilizando el método de *prompting* COD-RE. El modelo que presentó mejores resultados en todas las columnas fue *Qwen 3 8B fine-tuning* con excepción de la métrica *Precision*, donde resultó levemente superado por *Llama 3.1 Instruct fine-tuning*. En todos los modelos que contaron con pruebas de *fine-tuning* se observó un aumento remarcable en *Precisión*, siendo el más destacable dentro las tres métricas de BERTScore presentadas el modelo *Llama 3.1 Instruct* donde se obtuvo una diferencia de 0.0744 puntos, es decir, del 12.66% relativo. Cabe aclarar que no es posible obtener las métricas de los modelos base sin el *fine-tuning*. Por otro lado, el modelo Llama 3.1 8B Instruct COD-RE y el modelo Qwen 2.5 7B Instruct contaron con una diferencia de 0.0549 y 0.0349 puntos para ROUGE L, en otras palabras, un 63.91% y 41.01% relativo, respectivamente.

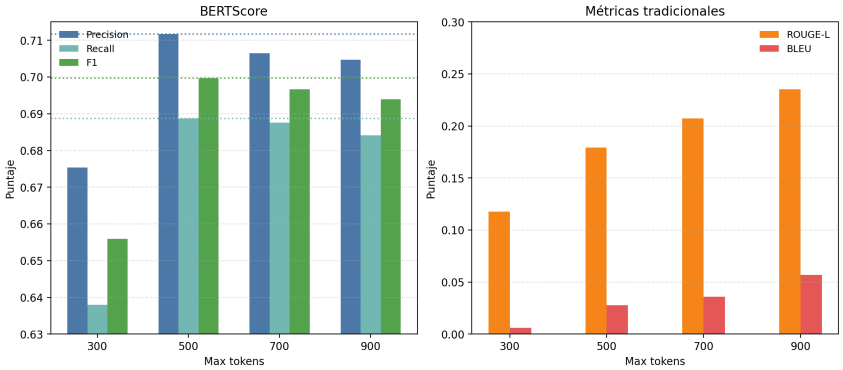


**Fig. 2.** Valores de BERTScore F1, ROUGE-L y BLEU en función del parámetro del número máximo de tokens.

Finalmente, el cuarto conjunto de experimentos se concentra en medir el impacto de las métricas al número máximo de tokens permitidos en el resumen. En figura 2 se presentan valores de las métricas utilizadas midiendo el impacto del máximo número de tokens para los valores 300, 500, 700 y 1000 *tokens* sobre el modelo Llama Instruct post *fine-tuning* (utilizando una muestra de 14612). Se percibe una relación exponencial sobre el aumento de los tokens durante la generación del modelo y las métricas tradicionales, a partir de este punto se presunta una estabilización de los resultados, mientras que, para BERTScore, se alcanzo un punto álgido en los 500 *tokens*. Esto podría explicar por qué métodos sin *tokens* regulados (PB y COD) obtienen mayores *scores* en la tabla 3, indicando que un resúmen más extenso esta directamente relacionado con mejores métricas ROUGE, BLEU Y BERTScore, sin embargo se consta que mayor longitud no tiene por que garantizar consecuentemente un resúmen de mayor calidad. Lo mismo ocurre en la tabla 3 para el modelo LLama 3.1 8B Instruct con método COD donde todas las métricas BERTScore caen luego de implementar COD-RE, por otro lado para el modelo Qwen los valores se mantuvieron estables a pesar de las variaciones entre COD y COD-RE.

## 4 Conclusión

La integración de resúmenes precisos en la metadata de los corpus jurídicos es fundamental para la indexación y recuperación eficiente de información. Mientras que tradicionalmente esta tarea dependía exclusivamente del análisis manual de profesionales del derecho, el avance de los grandes modelos de lenguaje (*LLMs*) ofrece hoy una oportunidad sin precedentes para sistematizar este proceso. Este trabajo demuestra que una metodología basada en modelos de código abierto, optimizada mediante *fine-tuning* con un corpus de jurisprudencia argentina altamente especializado —donde el 70% de los documentos contiene un resúmen por expertos—, permite automatizar la generación de síntesis con una alta fidel-

idad semántica. En última instancia, este enfoque preserva el rigor técnico del dominio legal y escala la capacidad de procesamiento de los documentos legales.

Se comprueba que el ajuste de los modelos mediante *fine-tuning* mejoran la performance en las respuestas tanto para métricas tradicionales, 63.91% relativo para ROUGE L en el modelo Llama 3.1 8B Instruct, como para BERTScore F1, 12.66% relativo para el mismo modelo.

El método de *prompting* COD-RE propuesto resulta efectivo para controlar la longitud de los *tokens* bajando en su totalidad la taza de resúmenes inválidos. Además, se realizaron experimentos que prueban que el tamaño del resúmen esta directamente relacionado con el resultado de las métricas ROUGE, BLUE y BERTScore, siendo que, a mayor número de *tokens* se obtienen mejores *scores* para métricas tradicionales mientras que para BERTScore se alcanzo la mayor performance en los 700 *tokens*.

Los resultados indican que la temperatura no juega un papel importante en el resultado de las métricas para modelos con *fine-tuning* pero resulta de relevancia para modelos no entrenados.

En trabajos futuros se planifica utilizar métricas de coherencia factual automatizadas, númericas y basadas en entidades, para medir el porcentaje de alucinación de las respuestas. Además, se plantea la evaluación de los resúmenes generados a través de aval y la retroalimentación de profesionales del campo o su imitación utilizando *LLMs*, como se realiza en Steffes et al., 2025, valorando de esta manera aspectos como la taza de alucinaciones, facilidad de lectura del resúmen o preferencia dependiendo del tipo de modelo utilizado. Finalmente, los resúmenes generados se integrarán en un sistema RAG formando parte de la metadata de su base de datos en conjunto con la metodología propuesta en Smulever et al., 2026.

## 5 Agradecimientos

LBC expresa un especial agradecimiento al programa de becas IA-UNNE (Eje Legal) de la Universidad Nacional del Nordeste, que nos brinda la oportunidad y financiación para iniciar nuestra formación en la investigación científica en el área de IA. Este trabajo utilizó recursos computacionales provistos por el Centro de Cómputos de Alto Desempeño del Nordeste Argentino (CECONEA) [9]. Agradecemos a *IJ International Legal Group* [10] por proveer la base de datos necesaria para el desarrollo de este trabajo.

## References

Adams, G., Fabbri, A., Ladhak, F., Lehman, E., & Elhadad, N. (2023). From sparse to dense: Gpt-4 summarization with chain of density prompting.

[9] http://cad.unne.edu.ar/index.php
[10] https://ij-ilg.com/

Ajay Mukund, S., & Easwarakumar, K. S. (2025). Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation. Symmetry, 17(5).

Akter, M., Çano, E., Weber, E., Dobler, D., & Habernal, I. (2025). A comprehensive survey on legal summarization: Challenges and future directions.

de Martim, H. (2025). An ontology-driven graph rag for legal norms: A structural, temporal, and deterministic approach.

FACENA. (2025). Ceconea.

Group, I. I. L. (2026). Ij international legal group.

Han, D., & team, U. (2023). Unsloth.

Heddaya, M., MacMillan, K., Malani, A., Mei, H., & Tan, C. (2024). Casesumm: A large-scale dataset for long-context summarization from u.s. supreme court opinions.

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.

Lin, C.-Y. (2004). A package for automatic evaluation of summaries. Text Summarization Branches Out, 74–81.

Meta-AI. (2024a). Llama-3.1-8b.

Meta-AI. (2024b). Llama-3.1-8b-instruct.

Meta-AI. (2024c). Meta-llama-3.1-8b-bnb-4bit.

Ortman, S., Canteros, L., Vargas, F., Escalante, G., González Coene, A., & Pulido, M. (2025). Anonimización de documentos legales usando llms con preentrenamiento continuo y sintonizado fino. JAIIO, Jornadas Argentinas de Informática, 11(1), 325–339.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 311–318.

Qwen-Team. (2024a). Qwen2.5-7b.

Qwen-Team. (2024b, September). Qwen2.5: A party of foundation models.

Ragazzi, L., Moro, G., Guidi, S., & Frisoni, G. (2024). Lawsuit: A large expert-written summarization dataset of italian constitutional court verdicts: Lawsuit: A large expert-written summarization dataset of... Artif. Intell. Law, 33(4), 1151–1187.

Richardson, L. (2007). Beautiful soup documentation. April.

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 379–389). Association for Computational Linguistics.

Smulever, F., Pulido, M., Canteros, L. B., & Mariño, S. I. (2026). Filtrado estructurado por metadatos como estrategia de refinamiento de la búsqueda semántica en sistemas rag para el dominio legal. JAIIO, Jornadas Argentinas de Informática.

Steffes, B., Rataj, P., Burger, L., & Roth, L. (2023). On evaluating legal summaries with rouge. Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, 457–461.

Steffes, B., Wiedemann, N. T., Gratz, A., Hochreither, P., Meyer, J. E., & Schilke, K. L. (2025). Summarisation of german judgments in conjunction with a class-based evaluation.

Team, Q. (2025, April). Qwen3.

Vargas, F., Coene, A. G., Escalante, G., Lobón, E., & Pulido, M. (2025). The impact of fine tuning in llama on hallucinations for named entity extraction in legal documentation.

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.

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).

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. International Conference on Learning Representations.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., & Chi, E. (2023). Least-to-most prompting enables complex reasoning in large language models.