Structured Metadata Filtering as a Semantic Search Refinement Strategy in RAG Systems for the Legal Domain

Federico Smulever[1], Manuel Pulido[1], Sonia I. Mariño[1], and Luciana B. Canteros[1]

FaCENA, Universidad Nacional del Nordeste, Corrientes, Argentina
fsmulever@comunidad.unne.edu.ar

Abstract. Retrieval-Augmented Generation (RAG) is a methodology that reduces hallucinations in language models applied to specialized databases. The standard RAG technique relies on semantic search to identify relevant documents for a user query, which are then integrated into the model's context to generate a response. In the legal domain, pure semantic search presents structural limitations in heterogeneous legal corpora: it operates over the entire vector space without prior discrimination, retrieving semantically similar but legally irrelevant documents. This paper proposes and empirically evaluates self-query as a semantic search refinement strategy in RAG systems applied to the Argentine legal domain. Self-query extends RAG by incorporating a stage that extracts legal labels from the user query, followed by semantic linking of the labels with the structured metadata of the documents in the corpus and filtering of the pertinent documents. Then, semantic search is applied to the filtered documents. The system was evaluated on a corpus of 26,998 Argentine legal documents with 32 structured metadata fields, through 684 independent evaluations performed by three language models as judges. Results show that self-query outperforms pure semantic search in 59.5 % of cases, with a consolidated average of 3.61 against 3.17 out of 5.00 and an inter-evaluator agreement of 90.8 %. The results show that structured metadata filtering constitutes an effective method of specific refinement of semantic search in legal domains with rich and heterogeneous taxonomies.

Keywords: RAG, self-query, Metadata filtering, Legal AI, LLM

Título: Filtrado Estructurado por Metadatos como Estrategia de Refinamiento de la Búsqueda Semántica en Sistemas RAG para el Dominio Legal

Resumen: La Generación Aumentada por Recuperación (RAG) es una metodología que permite reducir las alucinaciones en modelos de lenguaje aplicados a bases de datos especializadas. La técnica RAG estándar identifica documentos relevantes ante una consulta del usuario mediante búsqueda semántica, integrándolos al contexto del modelo para generar la respuesta. En el dominio legal, la búsqueda semántica presenta limitaciones en corpus heterogéneos: opera sobre la totalidad del espacio vectorial sin discriminación previa, recuperando

documentos semánticamente similares pero jurídicamente irrelevantes. Este trabajo propone y evalúa empíricamente self-query como estrategia de refinamiento de la búsqueda semántica en sistemas RAG para el dominio legal argentino. Self-query extiende el RAG incorporando una etapa de extracción de etiquetas legales desde la consulta en lenguaje natural, seguida de vinculación semántica con la metadata estructurada de los documentos, filtrando aquellos compatibles con los criterios identificados — logrando una reducción promedio del 98.50 % del conjunto de documentos candidatos — sobre los cuales se aplica la búsqueda semántica. El sistema se evaluó mediante un corpus de 26.998 documentos jurídicos argentinos con 32 campos de metadatos estructurados. Mediante 684 evaluaciones independientes realizadas por tres modelos de lenguaje (LLM-as-a-judge), los resultados demuestran que self-query supera a la búsqueda semántica pura en el 59.5 % de los casos, con un promedio consolidado de 3.61 frente a 3.17 sobre 5.00 y una concordancia inter-evaluador del 90.8 %, validando la integración de metadatos estructurados como mecanismo efectivo de filtrado y refinamiento semántico en entornos legales con taxonomías ricas y heterogéneas.

Palabras clave: RAG, self-query, Filtrado por metadata, IA Legal, LLM

1. Introducción

El desarrollo de los Modelos de Lenguaje de Gran Escala (LLMs, por sus siglas en inglés) ha transformado profundamente el procesamiento del lenguaje natural, habilitando capacidades avanzadas de comprensión y generación de texto en múltiples dominios. Sin embargo, su aplicación directa en áreas especializadas presenta limitaciones críticas, siendo la más significativa la generación de información no fundamentada en fuentes verificables, fenómeno conocido como alucinaciones (Dahl et al., 2024). En el ámbito jurídico, esta limitación adquiere especial relevancia: la precisión normativa y la trazabilidad de fuentes son requisitos indispensables para el ejercicio profesional del derecho. A su vez, el uso de LLMs en la práctica jurídica crece sostenidamente, con profesionales del derecho que incorporan estas herramientas en tareas de investigación, redacción de escritos y consulta normativa. Esta tendencia acentúa la necesidad de desarrollar sistemas que garanticen respuestas fundamentadas en documentación jurídica verificable.

La Generación Aumentada por Recuperación (RAG) (Lewis et al., 2020) surge como respuesta a estas limitaciones, combinando la capacidad generativa de los LLMs con un sistema de recuperación de documentos relevantes del corpus. En su forma más básica, RAG recupera documentos mediante búsqueda semántica por similitud vectorial: la consulta del usuario se transforma en un embedding y se compara contra los vectores que representan cada documento de la base de datos, para recuperar los más cercanos semánticamente. Si bien este enfoque mejora sustancialmente la fundamentación de las respuestas respecto a los LLMs puros, presenta una limitación estructural en corpus de gran escala con metadata rica: opera sobre la totalidad del espacio vectorial sin discriminación estructural previa (Gao et al., 2023). En un corpus jurídico heterogéneo,

esto implica que una consulta sobre derecho laboral puede recuperar documentos de derecho penal semánticamente relacionados pero jurídicamente irrelevantes, o que una consulta que requiere legislación específica retorne jurisprudencia o doctrina. La búsqueda semántica captura similitud de contenido pero ignora la estructura organizacional del corpus, generando falsos positivos que degradan la calidad de las respuestas generadas.

Los corpus legales modernos organizan sus documentos mediante metadata estructurada que incluye atributos como el tipo documental, la rama del derecho, el tribunal emisor y las figuras jurídicas abordadas, entre otros campos. Esta información no está capturada por los embeddings densos, que comprimen el significado semántico del texto pero abstraen su clasificación estructural. Trabajos previos han demostrado que el filtrado de base de datos mediante metadata extraída por LLMs mejora la selección de documentos relevantes en sistemas RAG (Poliakov y Shvai, 2024). Sin embargo, su aplicación y validación empírica en el dominio legal argentino, caracterizado por una taxonomía jurídica rica y heterogénea, permanece inexplorada. Este trabajo sostiene que utilizar la metadata estructurada del corpus como filtro previo a la búsqueda vectorial permite reducir drásticamente el espacio de búsqueda, eliminando documentos estructuralmente inadecuados antes de calcular similitud semántica.

En este trabajo se implementa y evalúa la técnica self-query sobre un corpus de documentos legales argentinos, comparándola empíricamente contra búsqueda semántica pura mediante un sistema de evaluación basado en múltiples modelos de lenguaje como jueces independientes (L. Zheng et al., 2023). Self-query extiende el pipeline RAG incorporando una etapa de extracción de etiquetas legales desde la consulta en lenguaje natural, donde un modelo de lenguaje determina qué campos de metadata son relevantes para la consulta y genera las etiquetas correspondientes. Estas son luego vinculadas con la metadata estructurada del corpus y aplicadas como filtros de forma jerárquica — desde criterios generales como el tipo documental hasta referencias específicas como la carátula de una jurisprudencia, el número de una norma o el autor de una doctrina —, reduciendo progresivamente el espacio de búsqueda antes de ejecutar la similitud vectorial. Los resultados demuestran que el sistema logra una reducción promedio del 98.50 % del conjunto de documentos candidatos mediante el filtrado estructurado por metadata, mejorando significativamente la relevancia de los documentos recuperados: self-query supera a la búsqueda semántica pura en el 59.5 % de los casos, validando que la información organizacional del corpus constituye un mecanismo efectivo de filtrado y refinamiento semántico en dominios jurídicos con taxonomías ricas y heterogéneas.

2. Trabajos Relacionados

La adaptación de LLMs al dominio jurídico ha sido objeto de investigación creciente en los últimos años. Chalkidis et al. (2020) desarrollaron LEGAL-BERT, un modelo pre-entrenado específicamente sobre textos legales que demostró mejoras significativas en tareas de clasificación y extracción de infor-

mación jurídica, estableciendo un precedente sobre la necesidad de especializar modelos para el dominio legal. En esta misma línea, Zheng et al. (2021) analizaron el impacto del preentrenamiento en tareas legales, y Yu, Quartey y Schilder (2022) exploraron estrategias de prompting para orientar LLMs hacia el razonamiento jurídico. Sin embargo, estos enfoques basados en modelos especializados o prompting presentan limitaciones persistentes ante consultas que requieren normativa o jurisprudencia específica y actualizada, donde las alucinaciones siguen siendo un problema documentado (Dahl et al., 2024).

La Generación Aumentada por Recuperación surge como respuesta a esta limitación, fundamentando las respuestas generadas en documentos recuperados de un corpus verificable (Lewis et al., 2020). Gao et al. (2023) documentaron en un survey exhaustivo las distintas arquitecturas RAG y sus casos de uso en dominios especializados, confirmando la viabilidad del enfoque para reducir alucinaciones. En esta línea, Asai et al. (2023) propusieron Self-RAG, un framework que extiende el pipeline RAG entrenando al modelo para decidir adaptativamente cuándo recuperar documentos y reflexionar sobre su propia generación, mejorando la factualidad sin sacrificar versatilidad. En el dominio legal específicamente, Wiratunga et al. (2024) propusieron CBR-RAG, combinando razonamiento basado en casos con recuperación aumentada para sistemas de pregunta-respuesta jurídica, demostrando mejoras en precisión respecto a RAG estándar. Estos trabajos convergen en señalar que la etapa de recuperación constituye un factor determinante en el desempeño de sistemas RAG legales, motivando el desarrollo de estrategias de recuperación más precisas.

A pesar de los avances mencionados, no se han reportado en la literatura trabajos que combine metadata estructurada como mecanismo de filtrado previo a la búsqueda vectorial en sistemas RAG sobre corpus legales. Los corpus jurídicos modernos disponen de clasificaciones temáticas, tipologías documentales y atributos institucionales que no están capturados por los embeddings densos pero que son determinantes para la relevancia jurídica de un documento. La heterogeneidad inherente del dominio legal hace que la similitud semántica pura sea una condición necesaria pero insuficiente para determinar relevancia. Esta brecha motiva el enfoque propuesto en el presente trabajo. En el contexto del corpus legal argentino utilizado, trabajos previos han abordado tareas de anonimización y extracción de entidades mediante LLMs (Ortman et al., 2025; Vargas et al., 2025), constituyendo antecedentes sobre el mismo conjunto de datos.

3. Corpus y Configuración Experimental

El corpus utilizado en este trabajo comprende 26.998 documentos jurídicos argentinos organizados en tres categorías documentales: Jurisprudencia (14.686 documentos, 54.4 %), Legislación (7.081 documentos, 26.2 %) y Doctrina (5.231 documentos, 19.4 %). La totalidad de los documentos corresponde a origen argentino, garantizando homogeneidad jurisdiccional y coherencia normativa. La base de datos fue provista por IJ International Legal Group (IJ-ILG), empresa

especializada en el desarrollo de herramientas jurídicas digitales para el ámbito legal argentino.

La heterogeneidad temática del corpus se refleja en 50 campos del derecho únicos con un promedio de 1.37 campos por documento, tales como Derecho Procesal, Derecho Penal, Derecho del Trabajo y Derecho Administrativo, y 1.493 voces jurídicas únicas con un promedio de 2.87 voces por documento, tales como Proceso Penal, Daño Moral, Medidas Cautelares o Defensa del Consumidor. La jurisprudencia se distribuye según la competencia del tribunal emisor: Tribunales Nacionales (6.559 fallos, 44.7 %), Tribunales Provinciales (4.433 fallos, 30.2 %) y Tribunales Federales (3.098 fallos, 21.1 %). Los documentos legislativos se clasifican por tipo normativo, incluyendo leyes, decretos y decretos de necesidad y urgencia, entre otros. Esta riqueza estructural de la metadata constituye el fundamento del enfoque propuesto: cada uno de estos atributos representa una dimensión de filtrado que la búsqueda semántica pura no puede explotar.

La distribución temporal del corpus abarca desde 1900 hasta 2024 en el caso de la legislación, desde 1980 hasta 2024 en jurisprudencia con concentración en la década 2010, y entre 2012 y 2024 en doctrina. Esta amplitud temporal implica que consultas sobre normativa vigente pueden recuperar documentos históricos semánticamente similares pero normativamente obsoletos, reforzando la necesidad de filtrado estructural previo.

Para la representación vectorial del corpus se utilizó el modelo BGE-M3 (Chen et al., 2024), que genera embeddings densos de 1.024 dimensiones con soporte multilingüe optimizado para español. Se adoptó una estrategia de segmentación (chunks) de 2.048 tokens con 20 % de solapamiento, configuración que demostró superioridad en evaluaciones preliminares sobre el corpus y que garantiza coherencia semántica en los fragmentos recuperados, generando un total de 147.217 chunks. Los vectores se almacenan en un índice FAISS compuesto por tres archivos complementarios que permiten, dado cualquier vector recuperado, identificar el documento fuente original del que proviene.

Para la evaluación experimental se diseñó un conjunto de 228 consultas legales elaboradas por profesionales del derecho, orientadas a cubrir la mayor diversidad posible del dominio jurídico argentino. Las consultas abarcan múltiples ramas del derecho — incluyendo Derecho Civil, Penal, Laboral y Administrativo, entre otras — y contemplan distintos niveles de especificidad: consultas conceptuales generales, consultas con restricciones temáticas explícitas, consultas de naturaleza procesal o de hecho, y consultas con referencias normativas o jurisprudenciales específicas. Esta diversidad garantiza que la evaluación comparativa refleje escenarios representativos del uso real del sistema en la práctica jurídica.

4. Arquitectura de self-query sobre Corpus Jurídico

4.1. Extracción de etiquetas legales

El primer componente del sistema self-query consiste en interpretar la consulta del usuario en lenguaje natural e identificar las etiquetas legales implícitas

o explícitas que contiene. Para esta tarea se empleó Llama 3.1 8B Instruct, un modelo de lenguaje cuantizado a 4 bits que opera sobre la consulta y produce como salida un objeto JSON con las etiquetas legales detectadas.

Las etiquetas extraíbles se organizan en cinco categorías según su naturaleza. Los atributos de tipo documental y temporal permiten identificar si la consulta refiere a jurisprudencia, legislación o doctrina, así como restricciones de fecha. La clasificación jurídica captura la rama del derecho y las voces jurídicas mencionadas implícita o explícitamente. Los atributos normativos permiten identificar números y tipos de norma cuando la consulta los menciona. Los atributos jurisdiccionales capturan el fuero y la clase de tribunal relevante para la consulta. Finalmente, las referencias jurisprudenciales específicas capturan el nombre o identificador de un fallo concreto cuando la consulta lo menciona explícitamente. Para cada categoría, el modelo asigna el valor extraído o null cuando no aplica a la consulta planteada. Ante extracciones incompletas en los campos de clasificación jurídica, el sistema ejecuta un segundo intento con prompt reformulado, mejorando la robustez ante variaciones en la formulación.

La evaluación de la calidad de este componente se presenta en la Sección 6.1.

4.2. Vinculación semántica de etiquetas con la metadata del corpus

Las etiquetas extraídas por el LLM expresan conceptos jurídicos en lenguaje natural que no necesariamente coinciden con los valores exactos almacenados en la metadata del corpus. Para garantizar la correspondencia entre etiquetas y metadata, el sistema aplica dos mecanismos según la naturaleza de la etiqueta.

Para etiquetas de naturaleza textual — tipo documental, rama del derecho, voces jurídicas, tipo de norma, clase y fuero de tribunal, y referencias jurisprudenciales — se aplica vinculación semántica mediante BGE-M3. El sistema proyecta tanto la etiqueta extraída como los valores únicos del campo correspondiente al espacio de embeddings, seleccionando el valor de referencia con mayor similitud de coseno junto con su score de confianza. Por ejemplo, una consulta puede generar la etiqueta “Corte Suprema” mientras que el valor de la metadata en el corpus es “Corte Suprema de Justicia de la Nación”. La vinculación semántica resuelve esta divergencia capturando correspondencias conceptuales más allá de la coincidencia literal. Para garantizar eficiencia computacional, los embeddings de los valores únicos de cada campo se calculan una única vez y se almacenan para su reutilización en consultas posteriores.

Para etiquetas de naturaleza numérica — números de leyes y decretos — se aplica un proceso de normalización de formato. El sistema estandariza la etiqueta extraída al formato exacto utilizado en la metadata del corpus: los números de ley se normalizan suprimiendo prefijos textuales (“Ley 20.744” $\rightarrow$ “20744”), y los números de decreto se normalizan al formato número/año (“Decreto 1234 de 2020” $\rightarrow$ “1234/2020”). Este proceso garantiza coincidencia exacta entre la etiqueta normalizada y el valor almacenado en la metadata, sin necesidad de búsqueda aproximada.

Como resultado de ambos procesos, el sistema obtiene un conjunto de filtros expresados en los valores precisos del corpus, listos para ser aplicados en la etapa de filtrado progresivo.

4.3. Filtrado progresivo de vectores del corpus jurídico

Con las etiquetas legales extraídas de la consulta, el sistema intenta vincularlas con los valores de metadata del corpus. Según el resultado de esta vinculación y la naturaleza de las etiquetas, se aplican dos estrategias de filtrado sobre el índice de 147.217 chunks.

Cuando la consulta contiene referencias específicas — como el número de una ley o la carátula de una sentencia - y estas lograron vincularse con la metadata correspondiente, el sistema aplica exclusivamente ese filtro, dado que su naturaleza específica permite recuperar puntualmente el documento referenciado sin necesidad de criterios adicionales. Si la referencia específica no pudo vincularse — porque la norma o sentencia solicitada no existe en el corpus — el sistema descarta esa etiqueta y ejecuta la estrategia de filtrado progresivo con las demás etiquetas disponibles.

En ausencia de referencias específicas, o como mecanismo de respaldo cuando estas no pudieron vincularse, el sistema aplica los filtros en orden decreciente de generalidad: primero tipo documental, luego rama del derecho, luego voces jurídicas, y finalmente atributos institucionales. Este orden es deliberado: aplicar primero los filtros más específicos podría vaciar prematuramente el conjunto de documentos candidatos si algún criterio es demasiado restrictivo. Comenzar por los filtros más generales garantiza una reducción gradual y controlada, preservando siempre un subconjunto de documentos sobre el cual aplicar los criterios siguientes. Cuando un filtro reduce el resultado a cero documentos, el sistema lo descarta automáticamente manteniendo los demás activos, evitando que un único criterio excesivamente restrictivo invalide toda la búsqueda.

La Figura 1 ilustra el proceso completo sobre una consulta representativa. Partiendo de 147.217 chunks, la aplicación secuencial de filtros reduce progresivamente el espacio: tipo documental (72.789 documentos), rama del derecho (6.083), voces jurídicas (394), mientras que los filtros de tribunal y fecha son descartados por no producir coincidencias. El resultado final de 394 documentos representa una reducción del 99.73 % del espacio de búsqueda original, sobre el cual se ejecuta la similitud vectorial con BGE-M3 para recuperar los chunks más relevantes.

En un entorno operativo orientado a profesionales del derecho, cuando las etiquetas extraídas resultan insuficientes para reducir significativamente el espacio de búsqueda, el sistema podría incorporar un mecanismo de consulta iterativa que solicite al usuario información adicional — como la rama del derecho, el tipo de documento o el período temporal de interés — antes de ejecutar la similitud vectorial, mejorando así la precisión de la recuperación.

```
================================================================================
Consulta: Sentencias de la Corte Suprema sobre contratos después de 2015
Docstore original: 147,217 documentos
--------------------------------------------------------------------------------

[1/3] Extrayendo filtros con LLM...
Filtros extraídos:
  doc_type: jurisprudencia
  rel_campos: Derecho Civil
  rel_voces: Contratos
  court_type_txt: Corte Suprema
  doc_date: 2015

[2/3] Vinculando con metadata real...
Filtros vinculados:
  doc_type: JU (score: 0.461)
  rel_campos: Derecho Civil (score: 1.000)
  rel_voces: Contratos (score: 1.000)
  court_type_txt: Corte Suprema (score: 1.000)
  doc_date: 2015 (score: 1.000)

[3/3] Aplicando filtros al docstore...
  Inicio: 147,217 docs
  ✓ Filtro doc_type: 72,789 docs
  ✓ Filtro rel_campos: 6,083 docs
  ✓ Filtro rel_voces: 394 docs
  X Filtro court_type_txt: 0 docs → DESCARTADO (mantiene 394 docs)
  X Filtro doc_date: 0 docs → DESCARTADO (mantiene 394 docs)

  Filtros descartados: court_type_txt, doc_date

✓ Resultado final: 394 documentos
  Reducción: 146,823 docs eliminados
  Porcentaje restante: 0.27%
```

Figura 1. Ejemplo de filtrado progresivo sobre el corpus legal argentino. Consulta: "Sentencias de la Corte Suprema sobre contratos después de 2015".

5. Metodología de Evaluación

5.1. Evaluación de la extracción de etiquetas legales

La calidad del componente de extracción de etiquetas legales se evalúa mediante tres modelos de lenguaje como jueces independientes — Llama 3.1 8B Instruct, Qwen 2.5 7B Instruct y Mistral 7B Instruct v0.3 —, paradigma cuya validez ha sido establecida por L. Zheng et al. (2023), quienes demostraron que LLMs como evaluadores alcanzan niveles de concordancia con jueces humanos superiores al 80 %. Para cada consulta del conjunto descripto en la Sección 3, el sistema genera las etiquetas legales correspondientes mediante el proceso descripto en la Sección 4.1. Cada modelo evaluador recibe la consulta original junto con las etiquetas extraídas y asigna una puntuación de 1 a 5, donde 1 representa etiquetas incorrectas o irrelevantes y 5 representa etiquetas completamente coincidentes con la naturaleza jurídica de la consulta. El promedio de las puntuaciones de cada modelo constituye su score individual, y el promedio consolidado final se calcula a partir de los promedios individuales de los tres evaluadores.

5.2. Evaluación comparativa: self-query vs. búsqueda semántica

Para cada consulta del conjunto descripto en la Sección 3, ambos enfoques recuperan de forma independiente los tres segmentos más relevantes del corpus. Los resultados se presentan a tres modelos de lenguaje como jueces independientes — Llama 3.1 8B Instruct, Qwen 2.5 7B Instruct y Mistral 7B Instruct v0.3 —

mediante un prompt en formato JSON que organiza los segmentos en Conjunto A y Conjunto B sin revelar qué técnica produjo cada uno, reduciendo posibles sesgos (L. Zheng et al., 2023). Cada evaluador asigna una puntuación del 1 al 5 a cada segmento, donde 1 indica documento irrelevante y 5 indica que responde directamente a la consulta. El enfoque con mayor promedio de puntuaciones es declarado ganador, registrándose empate en caso de puntuaciones idénticas. Los resultados se analizan en términos de distribución de victorias, promedio general de puntuaciones y concordancia entre evaluadores.

6. Resultados y Discusión

6.1. Evaluación de la extracción, vinculación y aplicación de las etiquetas legales

Siguiendo la metodología descripta en la Sección 5.1, se presentan los resultados de la evaluación del componente de extracción de etiquetas legales. La Tabla 1 reporta la calidad de las etiquetas extraídas según la valoración de los tres modelos jueces.

Cuadro 1. Evaluación cualitativa por modelo. Calidad de etiquetas.

Modelo	**Promedio general**	**Mediana**	**Desviación estándar**
Llama 3.1 8B	3.8312 / 5	4.0000	0.4439
Qwen 2.5 7B	4.1240 / 5	4.1240	0.9231
Mistral 7B V0.3	3.9485 / 5	3.9485	0.6645
Total	**3.9485 / 5**	–	–

Los tres modelos evaluadores mostraron puntuaciones consistentemente altas: 3.83/5.00 (Llama 3.1 8B), 4.12/5.00 (Qwen 2.5 7B) y 3.95/5.00 (Mistral 7B v0.3), con medianas de 4.00 en los tres casos y desviaciones estándar bajas, confirmando un promedio consolidado superior a 3.9/5.00. Para el análisis de concordancia, las puntuaciones de cada evaluador fueron categorizadas en tres rangos: Baja (1-2), Media (3) y Alta (4-5). Se consideró concordancia unánime cuando los tres modelos asignaron la misma categoría, concordancia mayoritaria cuando dos de los tres coincidieron, y discordancia total cuando los tres asignaron categorías diferentes. Sobre las 228 preguntas evaluadas, se observó concordancia unánime en 177 casos (77.6 %), concordancia mayoritaria en 47 casos (20.6 %), y discordancia total en apenas 4 casos (1.8 %). La concordancia total alcanzó el 98.2 %, evidenciando que los tres modelos independientes convergen en sus criterios de evaluación y validan la consistencia del proceso de extracción de etiquetas legales sobre el corpus jurídico argentino.

El análisis de las etiquetas extraídas sobre las 228 consultas evidencia una cobertura robusta del componente de extracción. En promedio se extrajeron

2.21 etiquetas por consulta (mediana 2), con un mínimo de 1 y un máximo de 5 etiquetas, sin registrarse ninguna consulta sin etiquetas. La distribución muestra que el 1.8 % de las consultas generó una etiqueta, el 83.8 % dos, el 9.2 % tres, el 3.1 % cuatro y el 2.2 % cinco, reflejando que el sistema adapta la extracción a la complejidad de cada consulta.

Los atributos de clasificación jurídica — rama del derecho (98.2 %) y voces jurídicas (97.4 %) — dominaron la extracción, resultado esperable dado que constituyen la dimensión más general para segmentar una base jurídica heterogénea: toda consulta legal tiene implícitamente asociada una rama del derecho y una figura jurídica. La alta presencia de estos atributos valida además el mecanismo de reintento descripto en la Sección 4.1, que garantiza una segunda oportunidad de extracción cuando no se logra capturar estos campos en la primera instancia. Los atributos más específicos — tipo documental (11.4 %), tipo de norma (4.8 %), fuero (3.5 %), clase de tribunal (2.6 %), número de norma (2.6 %) y carátula de documento (0.9 %) — aparecieron con menor frecuencia, correspondiendo a consultas con referencias institucionales o normativas explícitas, lo que confirma que el sistema discrimina correctamente entre consultas generales y específicas.

Complementariamente, se analizó la capacidad de vinculación de las etiquetas extraídas con la metadata de la base de datos. Sobre un total de 505 etiquetas extraídas, el 98.0 % (495) logró vincularse exitosamente con valores de metadata del corpus, mientras que apenas el 2.0 % (10) no pudo vincularse. El análisis de las etiquetas no vinculadas revela que en todos los casos corresponden a referencias normativas o jurisprudenciales específicas — leyes, decretos y sentencias — cuyos documentos no se encuentran en el corpus, confirmando que las fallas de vinculación responden a ausencias en la base de datos y no a errores del componente de extracción o vinculación semántica.

Como resultado de la vinculación exitosa de las etiquetas con la metadata del corpus, la reducción del conjunto de chunks sobre el cual opera la búsqueda semántica fue sustancial. Sobre las 228 consultas evaluadas, la aplicación de los filtros redujo en promedio el 98.50 % de los chunks disponibles (mediana 99.88 %), con un mínimo de 50.56 % y un máximo del 100 %. En todos los casos, la búsqueda semántica operó exclusivamente sobre el subconjunto filtrado, con una media de 2.204 chunks de los 147.217 disponibles, garantizando que la similitud vectorial se calcule únicamente sobre documentos estructuralmente compatibles con la consulta planteada.

6.2. Comparación self-query vs. Búsqueda semántica

Los resultados de la comparación entre self-query y búsqueda semántica pura se presentan en la Tabla 2 y la Tabla 3. Las métricas integran un total de 684 evaluaciones independientes, producto de someter 228 consultas al criterio de tres modelos jueces.

Los tres modelos evaluadores mostraron preferencia consistente por self-query, con victorias que oscilan entre el 56.1 % (Qwen 2.5 7B) y el 64.0 % (Mistral 7B V0.3), frente a un máximo del 31.6 % para búsqueda semántica.

Cuadro 2. Evaluación cualitativa por modelo. Self-query vs Búsqueda semántica.

Modelo	Métrica	Self-query	Búsqueda semántica	Empates
Llama 3.1 8B	Victorias	133 (58.3 %)	69 (30.3 %)	18 (7.9 %)
	Promedio	3.54 / 5	2.83 / 5	–
Qwen 2.5 7B	Victorias	128 (56.1 %)	72 (31.6 %)	28 (12.3 %)
	Promedio	3.57 / 5	3.28 / 5	–
Mistral 7B V0.3	Victorias	146 (64.0 %)	58 (25.4 %)	24 (10.5 %)
	Promedio	3.73 / 5	3.40 / 5	–

El análisis consolidado confirma esta tendencia: 407 victorias para self-query (59.5 %) frente a 199 para búsqueda semántica (29.1 %), con promedios consolidados de 3.61 y 3.17 respectivamente, representando una mejora de 0.44 puntos. La consistencia de los resultados entre tres modelos de distintas arquitecturas refuerza la validez de los hallazgos, descartando que la preferencia por self-query sea un artefacto de un evaluador particular.

Cuadro 3. Evaluación consolidada. Self-query vs Búsqueda semántica.

Métrica	Self-query	Búsqueda semántica	Empates
Victorias totales	407 (59.5 %)	199 (29.1 %)	70 (10.2 %)
Promedio general	3.61 / 5	3.17 / 5	–

El análisis de concordancia revela que en el 55.3 % de los casos los tres evaluadores coincidieron unánimemente en su decisión, y en el 35.5 % adicional dos de tres coincidieron, alcanzando una concordancia agregada del 90.8 %. Este alto nivel de acuerdo entre modelos de distintas arquitecturas refuerza la confiabilidad de los resultados, indicando que la preferencia por self-query no responde a sesgos propios de un evaluador particular sino que refleja una superioridad consistente y reproducible del enfoque.

Cuadro 4. Análisis de concordancia entre evaluadores.

Concordancia	Self-query	Búsqueda semántica	Empates
Unanimidad 126 (55.3 %)	86	33	7
Mayoría 81 (35.5 %)	55	26	–
Divididas 21 (9.2 %)		21 (casos sin consenso)	

Con el objetivo de caracterizar los casos en que la búsqueda semántica superó a self-query, se analizaron los 59 casos donde búsqueda semántica obtuvo victoria unánime o mayoritaria, comparando las etiquetas aplicadas por self-query

con la metadata de los documentos recuperados por búsqueda semántica. En el 61.0 % de estos casos la rama del derecho extraída no coincidía con la metadata relevante, y de los 22 casos restantes donde sí coincidía, el 81.8 % presentó divergencias en las voces jurídicas. En conjunto, el 93.2 % de las victorias de búsqueda semántica se explica por un filtrado estructural incorrecto producto de etiquetas inexactas, confirmando que la principal limitación de self-query radica en la calidad del proceso de extracción. Los 4 casos restantes constituyen excepciones aisladas — representando apenas el 1.8 % del total de consultas evaluadas — donde ambos enfoques extrajeron etiquetas correctas pero self-query recuperó documentos de menor calidad.

6.3. Discusión

Estos resultados validan la hipótesis central del trabajo: el filtrado estructurado por metadata reduce drásticamente el conjunto de chunks candidatos — con una reducción promedio del 98.50 % y una mediana del 99.88 % sobre las 228 consultas evaluadas — eliminando documentos que, si bien son semánticamente similares a la consulta, son estructuralmente incompatibles con ella. Una consulta sobre derecho laboral puede recuperar documentos de derecho penal con alta similitud coseno, pero self-query los excluye antes de ejecutar la búsqueda vectorial al circunscribir el espacio a la rama del derecho correcta. Esta discriminación estructural previa es precisamente lo que la búsqueda semántica pura no puede realizar, y los resultados cuantifican su impacto: una mejora de 0.44 puntos en relevancia promedio y una proporción de victorias considerables en favor de self-query.

No obstante, los resultados también revelan que la principal limitación del enfoque radica en la calidad del proceso de extracción de etiquetas legales. El análisis de los 59 casos donde búsqueda semántica obtuvo victoria unánime o mayoritaria demostró que el 93.2 % se explica por etiquetas incorrectas que derivaron en una vinculación errónea con la metadata del corpus — o directamente en la imposibilidad de vincularlas cuando no tienen representación en la base de datos —, generando un filtrado incorrecto que circunscribió la búsqueda a documentos irrelevantes para la consulta. Este riesgo no existe en la búsqueda semántica pura, que opera directamente sobre la representación vectorial de la consulta sin intermediación de un proceso de extracción susceptible a errores. Esto sugiere que self-query refina la búsqueda semántica cuando las etiquetas extraídas son precisas, pero su efectividad está condicionada por la calidad del componente de extracción, lo que motiva el desarrollo de mecanismos de validación y corrección de etiquetas como línea de trabajo futuro.

7. Conclusiones

Este trabajo demostró empíricamente que el filtrado estructurado por metadata mejora significativamente la relevancia de los documentos recuperados

en sistemas RAG aplicados al dominio legal argentino. Los resultados obtenidos sobre 684 evaluaciones independientes confirman que self-query supera a la búsqueda semántica pura en el 59.5 % de los casos, con un promedio consolidado de 3.61 vs. 3.17 sobre 5.00, validando la hipótesis central del trabajo: la información organizacional del corpus jurídico constituye un mecanismo de refinamiento de la búsqueda semántica que la similitud vectorial pura no puede explotar.

La superioridad de self-query se explica por la reducción drástica del conjunto de vectores candidatos previo a la búsqueda semántica, con una reducción promedio del 98.50 % y una mediana del 99.88 % sobre las 228 consultas evaluadas. Al circunscribir la búsqueda a subconjuntos estructuralmente homogéneos del corpus, el sistema elimina documentos que son jurídicamente irrelevantes. Esta discriminación estructural previa es la que explica la mejora de 0.44 puntos en relevancia promedio y la proporción de victorias de 2:1 observada en los resultados.

No obstante, el análisis de los casos donde búsqueda semántica superó a self-query revela que la principal limitación operativa del enfoque radica en la calidad del proceso de extracción de etiquetas: el 93.2 % de dichos casos se explica por etiquetas incorrectas que derivaron en una vinculación errónea con la metadata del corpus, generando un filtrado incorrecto que circunscribió la búsqueda a documentos irrelevantes para la consulta, limitación ausente en la búsqueda semántica pura. A su vez, self-query presenta una limitación estructural inherente al paradigma RAG tradicional: cada documento es tratado como una unidad independiente, ignorando las interrelaciones constitutivas del razonamiento jurídico — normas que derogan otras normas, jurisprudencias fundamentadas en normativa específica y doctrina que analiza fallos y leyes.

Como trabajo futuro orientado a mejorar self-query, se propone incorporar un mecanismo de consulta iterativa que ante etiquetas insuficientes solicite información adicional al usuario — siguiendo enfoques recientes de políticas de clarificación adaptativas (Berant et al., 2025) —, explorar mecanismos de validación automática de etiquetas legales para mitigar la principal limitación identificada, y evaluar el impacto de modelos de embeddings más avanzados sobre la calidad de la búsqueda semántica aplicada al subconjunto filtrado. A más largo plazo, superar la independencia estructural entre documentos propia del paradigma RAG tradicional requiere modelar el corpus legal como un grafo de conocimiento que capture las relaciones explícitas entre normas, fallos y doctrina, constituyendo una línea de investigación futura de alto potencial para el desarrollo de sistemas de recuperación jurídica de mayor precisión.

Agradecimientos

Los autores agradecen a la Universidad Nacional del Nordeste por el marco institucional que hizo posible este trabajo, a las becas IA-UNNE promovidas por la Secretaría General de Ciencia y Técnica por el apoyo al desarrollo académico y profesional, y a IJ International Legal Group (IJ-ILG) por el acompañamiento

en el proyecto y la provisión de la base de datos utilizada, en el marco de una iniciativa de investigación académica con vinculación tecnológica.

Referencias

Asai, Akari et al. (2023). «Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection». En: *arXiv preprint arXiv:2310.11511*.

Berant, Jonathan et al. (2025). «Learning Steerable Clarification Policies with Collaborative Self-play». En: *arXiv preprint arXiv:2512.04068*.

Chalkidis, Ilias et al. (2020). «LEGAL-BERT: The Muppets Straight out of Law School». En: *arXiv preprint arXiv:2010.02559*.

Chen, Jianlv et al. (2024). «BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation». En: *arXiv preprint arXiv:2309.07597*.

Dahl, Matthew et al. (2024). «Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models». En: *Journal of Empirical Legal Studies* 21.4, págs. 1-27.

Gao, Yunfan et al. (2023). «Retrieval-Augmented Generation for Large Language Models: A Survey». En: *arXiv preprint arXiv:2312.10997*.

Lewis, Patrick et al. (2020). «Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks». En: *Advances in Neural Information Processing Systems*. Vol. 33, págs. 9459-9474.

Ortman, Sofia Ornella et al. (2025). «Anonimización de Documentos Legales usando LLMs con Preentrenamiento Continuo y Sintonizado Fino». En: *Memorias del Simposio Argentino de Inteligencia Artificial y Ciencia de Datos (ASAID-JAIIO)*.

Poliakov, Mykhailo y Nadiya Shvai (2024). «Multi-Meta-RAG: Improving RAG for Multi-Hop Queries using Database Filtering with LLM-Extracted Metadata». En: *Proceedings of ICTERI 2024*. DOI: 10.1007/978-3-031-81372-6_25.

Vargas, Francisco et al. (2025). «The Impact of LLaMA Fine Tuning on Hallucinations for Name Entity Extraction in Legal Documents». En: *Electronic Journal of SADIO* 24.

Wiratunga, Nirmalie et al. (2024). «CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering». En: *International Conference on Case-Based Reasoning*, págs. 445-460.

Yu, Fangyi, Lee Quartey y Frank Schilder (2022). «Legal Prompting: Teaching a Language Model to Think Like a Lawyer». En: *arXiv preprint arXiv:2212.01326*.

Zheng, Lianmin et al. (2023). «Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena». En: *Advances in Neural Information Processing Systems*. Vol. 36.

Zheng et al. (2021). «When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset». En: *arXiv preprint arXiv:2104.08671*.