CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): [ ] Memoria o trabajo de título [X] Tesis de Postgrado

Título del trabajo: Implementación de búsqueda semántica y validación de jurisprudencia mediante LLM y RAG en la Corte de Apelaciones de Iquique.

Nombre del candidato(a): José Bitrán Filipp

Carrera / Grado: Magíster en Tecnologías de la Información

Campus: CASA CENTRAL **Departamento: INFORMÁTICA**

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, RICARDO ÑANCULEF, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

[X] El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

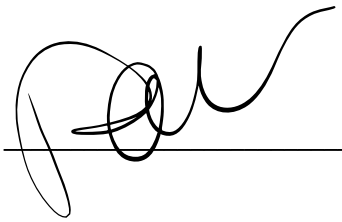
☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 20/04/1981 **Firma:**



Estudiante o Candidato(a):

Fecha: 12-03-2026 **Firma:**

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Implementación de búsqueda semántica y validación de jurisprudencia mediante LLM y RAG en la Corte de Apelaciones de Iquique

José Bitrán Filipp
jose.bitran@usm.cl

Resumen

La búsqueda y validación de jurisprudencia local en la Corte de Apelaciones de Iquique se realiza actualmente de forma mayoritariamente manual, lo que implica un alto consumo de tiempo en la identificación y análisis de información relevante, afectando directamente a los funcionarios de la unidad de secretaría de ministros y al proceso previo a la redacción de sentencias. Esta situación se ve acentuada por la ausencia de herramientas que permitan realizar búsquedas semánticas sobre las sentencias locales.

Con el objetivo de reducir el tiempo dedicado por ministros, relatores y funcionarios a la investigación y análisis previo a la redacción de sentencias, esta investigación exploró la aplicación de inteligencia artificial en el proceso judicial mediante el desarrollo de un Producto Mínimo Viable (MVP) basado en un modelo de lenguaje de gran tamaño (LLM) y técnicas de generación aumentada por recuperación (RAG). El sistema propuesto permitió realizar búsquedas semánticas sobre un corpus de sentencias laborales de la Corte de Apelaciones de Iquique, incorporando mecanismos de recuperación contextual y generación controlada de respuestas.

La evaluación del MVP se realizó mediante un enfoque híbrido que combinó métricas automáticas reference-free y validación experta. En términos de coherencia textual, la métrica BLANC mostró valores estables y consistentes, evidenciando un uso efectivo de los fragmentos de jurisprudencia recuperados. Asimismo, la calidad semántica de las respuestas, evaluada mediante la métrica G-EVAL con un comité de modelos de lenguaje, alcanzó promedios cercanos a 3,8 sobre 5. Por su parte, la validación experta realizada por funcionarios de la unidad de secretaría de ministros otorgó una evaluación global promedio de 4,7 sobre 5, confirmando la utilidad práctica del sistema en un contexto institucional real.

Adicionalmente, se aplicó una prueba de transferibilidad utilizando un corpus independiente de sentencias laborales de la Corte de Apelaciones de Arica, obteniéndose resultados comparables en coherencia, calidad y evaluación experta, sin degradaciones significativas respecto del corpus base. El análisis del acuerdo inter-juez mediante la métrica Kappa de Fleiss evidenció niveles de concordancia aceptables a moderados, reforzando la consistencia del proceso evaluativo.

Este trabajo no se orientó al desarrollo de nuevas arquitecturas de software, sino a evaluar la utilidad práctica de la implementación de inteligencia artificial en el contexto jurídico chileno. Los resultados obtenidos aportan evidencia empírica sobre la viabilidad, estabilidad y potencial utilidad de sistemas basados en IA como apoyo efectivo en la búsqueda y validación de jurisprudencia para la redacción de sentencias.

Palabras Clave: Inteligencia Artificial, Jurisprudencia, Búsqueda Semántica, RAG, Procesamiento de Lenguaje Natural, Modelos de Lenguaje (LLM)

1. Introducción

1.1. Contexto y Problemática

El proceso judicial tiene como objetivo final la dictación de justicia mediante la emisión de una decisión, la cual se materializa a través de una resolución de sentencia o fallo judicial. Si bien la redacción de estas sentencias es responsabilidad exclusiva de los ministros en el caso de la Corte de Apelaciones, dicha labor se apoya en funcionarios encargados de la búsqueda de sentencias previas, entendidas como resoluciones judiciales dictadas con anterioridad en casos similares, las cuales constituyen lo que se conoce como

jurisprudencia y permiten fundamentar jurídica y argumentativamente una nueva decisión. En este contexto, la búsqueda y validación de jurisprudencia constituye una actividad fundamental dentro del ámbito judicial, especialmente en los tribunales de alzada, donde se recepcionan recursos de diversas materias.

No obstante, la búsqueda y validación de jurisprudencia se realiza actualmente de forma completamente manual, mediante la revisión directa de carpetas compartidas que contienen fallos históricos en distintos formatos. Esta modalidad de trabajo demanda una cantidad considerable de tiempo, estimada entre 60 y 120 minutos por búsqueda, dependiendo de la complejidad del caso, según la experiencia reportada por los propios funcionarios involucrados (ver Anexo N°1, Anexo N°6), quienes deben revisar grandes volúmenes de documentos no indexados ni estandarizados para identificar información relevante que sirva de fundamento a nuevas sentencias.

Esta situación se ve agravada por la ausencia de un sistema eficiente que permita realizar búsquedas semánticas y contextualizadas sobre la jurisprudencia generada por la Corte de Apelaciones de Iquique. En definitiva, el tiempo invertido en la búsqueda de información se convierte en el principal problema del proceso, incidiendo directamente en la eficiencia operativa y restando tiempo a labores propias de la función jurisdiccional, tales como el análisis jurídico y la redacción de sentencias.

Según estadísticas oficiales de la Corte de Apelaciones de Iquique, durante el año 2024 se emitieron 5.885 sentencias judiciales en diversas materias, lo que evidencia una constante y dinámica generación de jurisprudencia. Este volumen de resoluciones constituye un insumo fundamental para la toma y redacción de futuras decisiones judiciales, especialmente en un tribunal de alzada donde la coherencia y consistencia jurisprudencial resultan esenciales. En el caso de la Corte de Apelaciones de Iquique, la jurisprudencia puede ser generada por ministros titulares, ministros suplentes y abogados integrantes, lo que incrementa la necesidad de contar con un sistema de acceso y búsqueda centralizado que permita disponer de la información de manera eficiente. En este contexto, la ausencia de herramientas refuerza la relevancia del problema abordado en esta investigación y justifica la exploración de soluciones apoyadas en inteligencia artificial para optimizar dicho proceso.

La figura 1 muestra el flujo actual del proceso de búsqueda de jurisprudencia, sus etapas y los participantes del proceso.

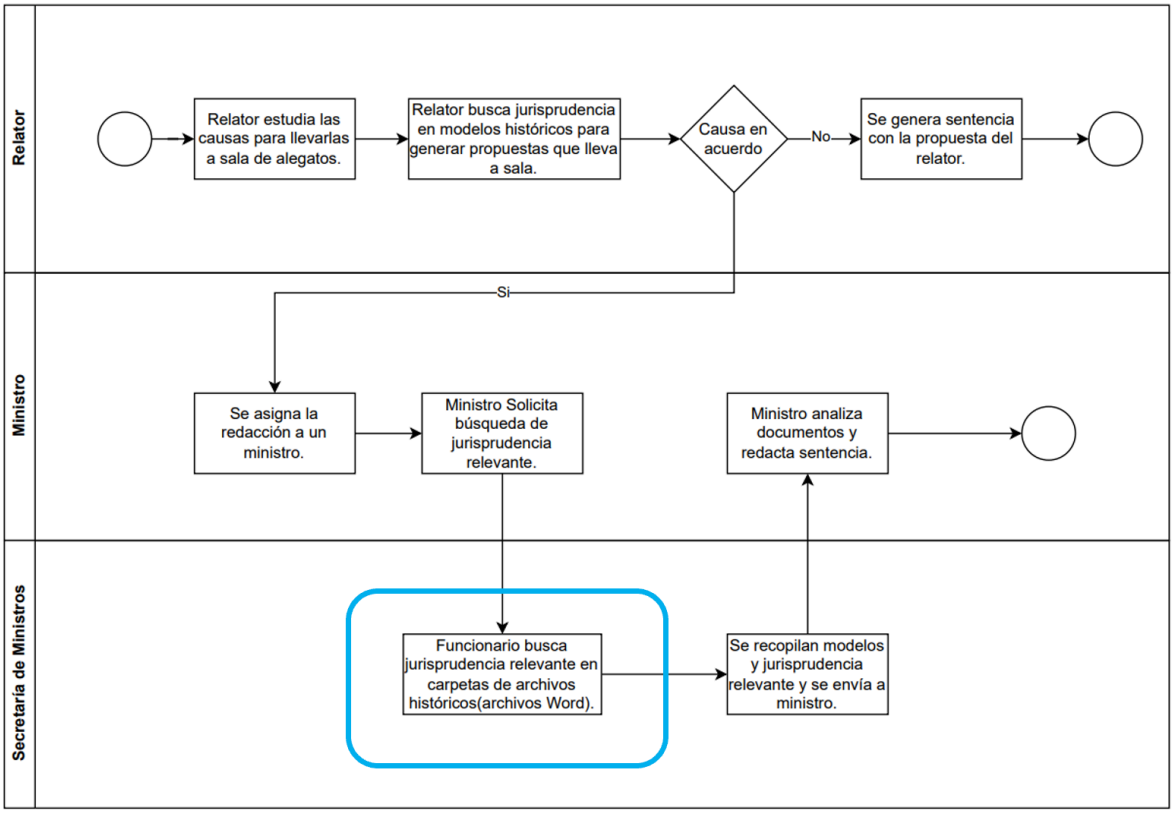


Figura1. Diagrama de actividades del proceso actual de búsqueda de jurisprudencia relevante.

Como se observa en el diagrama, el proceso actual depende principalmente de trabajo efectuado de forma manual por parte de relatores y funcionarios, lo que se traduce en demoras significativas en los tiempos de búsqueda de jurisprudencia, riesgos de omisión y falta de estandarización. Lo anteriormente descrito es lo que esta propuesta aborda en el capítulo 3 de este trabajo.

1.2. Propuesta de solución

Frente a la problemática identificada, se desarrolló una investigación aplicada que integró un análisis teórico sobre la inteligencia artificial aplicada al derecho, con énfasis en la búsqueda semántica y la validación de jurisprudencia. En este contexto, se ha evidenciado que la inteligencia artificial generativa, gracias a los avances en modelos de lenguaje de gran tamaño, ha permitido mejorar significativamente la comprensión y generación de texto en lenguaje natural, posibilitando su aplicación en escenarios complejos como el análisis jurídico y la recuperación de información jurisprudencial.
Como parte de esta solución, se diseñó e implementó un Producto Mínimo Viable (MVP) basado en un modelo de lenguaje de gran tamaño (LLM), técnicas de generación aumentada por recuperación (RAG), el framework LangChain y una base de datos vectorial (ChromaDB). El sistema desarrollado permitió realizar búsquedas semánticas de jurisprudencia en materia laboral, con el objetivo de reducir el tiempo que funcionarios, relatores y ministros dedican a la revisión manual de antecedentes, así como centralizar la información jurisprudencial en un repositorio único, eliminando la dependencia de archivos locales no estructurados.
La selección de las tecnologías utilizadas respondió a criterios de adecuación al procesamiento del español jurídico, capacidad de generación controlada basada en evidencia documental, y compatibilidad con arquitecturas de despliegue local. Asimismo, se priorizaron tecnologías de código abierto y configurables, con el objetivo de asegurar flexibilidad, control institucional y resguardo de la información jurídica tratada.

A continuación se muestran los principales usuarios del sistema.

Tabla 1. Caracterización del grupo objetivo

Rol	Funciones en el proceso actual	Nivel de interacción con el sistema propuesto
Funcionarios/as de la unidad de secretaría de ministros	Búsqueda de jurisprudencia, recopilación de fallos actuales y revisión de fondo y forma de estos.	**Alta:** Son los usuarios principales y participantes activos del desarrollo del MVP.
Ministros/as titulares.	Redacción de fallos, ocasionalmente realizan búsqueda de jurisprudencia.	**Media:** Usuarios recepcionan el insumo y se centran en la revisión y toma de decisión.
Relatores	Apoyo técnico durante las audiencias. Llevan propuestas de fallos a la sala de alegatos(acoge/rechaza), realizando para ello búsqueda de jurisprudencia.	**Alta:** Utilizan jurisprudencia para apoyar la decisión en sala.
Ministros suplentes y abogados integrantes	Participan en ciertas audiencias a petición del ministro presidente.	**Baja**: Son usuarios externos que reciben jurisprudencia y modelos de los funcionarios de la unidad de secretaría de ministros.

1.3. Hipótesis

Un sistema de búsqueda de jurisprudencia basado en RAG con búsqueda semántica genera respuestas de calidad suficiente para su uso en contextos judiciales reales, evidenciado por puntajes satisfactorios en métricas automáticas no supervisadas y una valoración positiva de expertos jurídicos mediante escala Likert, reduciendo además el tiempo reportado para esta tarea.

1.3.1. Variable Dependiente

Desempeño del sistema de búsqueda de jurisprudencia, medido en sus dimensiones de calidad, utilidad y eficiencia.

1.3.2. Variable Independiente

Sistema de IA especializado en búsqueda semántica de jurisprudencia.

Esta variable es la que impacta en la variable dependiente. Para este caso particular, es la utilización del MVP y cómo afecta el desempeño del sistema en sus dimensiones de calidad, utilidad y eficiencia en la búsqueda de jurisprudencia.

1.4. Objetivos planteados

Los objetivos planteados en este trabajo son los siguientes:

Objetivo General

Desarrollar un sistema basado en inteligencia artificial que permita reducir el tiempo requerido para la búsqueda semántica y validación de jurisprudencia en materia laboral, mediante un Producto Mínimo Viable (MVP), con el fin de apoyar las labores de los funcionarios de la Corte de Apelaciones de Iquique.

Objetivos Específicos

- Analizar los fundamentos teóricos y tecnológicos de la inteligencia artificial aplicada al derecho, con foco en la generación aumentada por recuperación (RAG) y en modelos de lenguaje de gran tamaño.
- Diseñar e implementar, en un entorno cloud, un Producto Mínimo Viable (MVP) que integre dichas tecnologías.
- Comparar el flujo de trabajo actual con el flujo propuesto, identificando las principales diferencias desde el punto de vista de la eficiencia.
- Evaluar el sistema propuesto mediante métricas automáticas y la evaluación de expertos jurídicos.

1.4.1. Metodología de trabajo

La metodología aplicada en esta investigación se estructuró en las siguientes etapas:

- Análisis teórico: Revisión de conceptos jurídicos y tecnológicos relacionados con inteligencia artificial.
- Construcción de Corpus: Recopilación y preparación de sentencias laborales de la Corte de Apelaciones de Iquique.
- Diseño e implementación de MVP: Desarrollo de un prototipo funcional integrando modelos de lenguaje de gran tamaño (LLM), técnicas RAG, el framework LangChain y una base de datos vectorial (ChromaDB).
- Comparación de flujos: Se contrasta el flujo de trabajo manual con el flujo propuesto.
- Evaluación del sistema: Evaluación del MVP mediante un enfoque híbrido que combina métricas automáticas (BLANC y G-EVAL) y validación experta basada en un cuestionario tipo Likert aplicado a usuarios jurídicos especializados

1.4.2. Marco teórico adoptado

El marco teórico que sustenta esta investigación se centra en definir los principales conceptos jurídicos implicados en la investigación, así como también, en la definición de las tecnologías utilizadas. Dichos conceptos en conjunto componen el contexto que orientó el diseño, desarrollo y evaluación del MVP.

1.4.3. Estrategia de Validación

Para evaluar la calidad y utilidad práctica del sistema propuesto, se utilizó un instrumento de evaluación basado en la escala Likert de cinco niveles (1 a 5), el cual fue aplicado a funcionarios clave de la unidad de secretaría de ministros. Las respuestas fueron evaluadas considerando cuatro criterios (Ver Anexo N°1): pertinencia jurisprudencial, entendida como la relevancia de las sentencias entregadas en relación con la consulta realizada; exactitud normativa y contextual, referida a la coherencia entre la descripción generada por el sistema y el contenido real de la sentencia; valor para fundamentar una decisión, asociado a la utilidad de la jurisprudencia recuperada para el sustento de nuevas resoluciones; y claridad y usabilidad, vinculada a la forma en que el sistema presenta la información, considerando elementos como rol, año y resumen de la sentencia. Cada consulta realizada al sistema fue evaluada en base a estos criterios, con el objetivo de cuantificar la calidad de las respuestas, estableciendo un puntaje máximo de 20 puntos por pregunta.
Complementariamente, para la validación de la hipótesis asociada a la reducción de tiempos de búsqueda y validación de jurisprudencia, se solicitó a los usuarios expertos que estimaran el tiempo promedio requerido para realizar dichas tareas mediante el proceso manual tradicional, así como el tiempo estimado al utilizar el sistema propuesto. Esta medición se basó en la experiencia operativa de los funcionarios participantes y se empleó como un indicador de impacto práctico del MVP en un contexto institucional real.
Adicionalmente, se incorporaron métricas automáticas con el fin de evaluar la calidad y coherencia de las respuestas generadas por el sistema desde una perspectiva objetiva. En particular, se utilizó la métrica BLANC, orientada a medir el grado en que la respuesta generada mejora la coherencia del texto recuperado, evaluando el aporte real del modelo generativo al contenido original. Asimismo, se empleó la métrica G-EVAL, basada en el uso de modelos de lenguaje como jueces automáticos, permitiendo evaluar dimensiones cualitativas de las respuestas, tales como coherencia, relevancia y consistencia argumentativa, en función de criterios previamente definidos.
Con el propósito de analizar el nivel de concordancia entre las evaluaciones realizadas por los expertos jurídicos, se aplicó el coeficiente Kappa de Fleiss, el cual permite medir el grado de acuerdo interevaluador más allá del azar, fortaleciendo la validez de los resultados obtenidos mediante evaluación humana. Finalmente, se consideró una prueba de transferibilidad, orientada a evaluar la capacidad del sistema para mantener un desempeño consistente frente a consultas distintas de aquellas utilizadas durante su configuración inicial, permitiendo analizar su potencial de generalización a nuevos escenarios de búsqueda jurisprudencial.

1.5. Estructura del informe

En el segundo capítulo se revisa el marco teórico de los conceptos utilizados en esta investigación y estado del arte. En el tercer capítulo, se aborda el desarrollo de la solución propuesta, los métodos de validación, los resultados obtenidos y el análisis de estos.
Finalmente, se presenta una conclusión respecto de los resultados, discusión de hipótesis, alcances limitaciones y trabajos futuros.

2. Marco Teórico y Estado del Arte

Este trabajo se apoya en conceptos del ámbito legal y la tecnología. Mediante la compresión de estos es posible abordar eficientemente la problemática, la hipótesis y la propuesta de solución. A continuación, se exponen los conceptos en mayor detalle, además del estado del arte relacionado con este trabajo.

2.1. Marco Teórico

2.1.1. Fallo Judicial y Jurisprudencia

La sentencia o fallo judicial es un acto jurídico procesal que dirime un conflicto, reconoce, declara o extingue una situación jurídica con implicaciones sociales directas, dictado por un representante de uno de los poderes del Estado, quien debe actuar conforme a los principios de legalidad, seguridad jurídica y respeto de los derechos fundamentales, dentro de un marco normativo establecido. La sentencia constituye, ante todo, un acto del juez o magistrado, lo que implica que cada resolución posee un sello y estilo argumentativo propio. Considerando el impacto social que generan las decisiones judiciales, resulta relevante la exigencia de celeridad, precisión y consistencia en su redacción. Asimismo, toda sentencia judicial debe ser emitida dentro de los plazos establecidos por la ley, los cuales varían según la materia de que se trate, siendo un ejemplo de ello el ámbito laboral, donde los fallos deben ser redactados dentro de un plazo de cinco días hábiles una vez que la causa ha quedado en estado de acuerdo en segunda instancia [1].
La jurisprudencia corresponde al conjunto de sentencias, decisiones o fallos dictados por los tribunales de justicia o por otras autoridades, en los casos en que la Constitución Política o la ley así lo faculta [2]. Estos antecedentes resultan fundamentales en la redacción de fallos judiciales, ya que permiten garantizar la consistencia y la seguridad jurídica, asegurando que situaciones similares reciban un tratamiento similar, lo que contribuye a generar confianza en el sistema judicial y a reducir la arbitrariedad en la toma de decisiones. Asimismo, la jurisprudencia se nutre de las experiencias de casos anteriores, actuando como un mecanismo de aprendizaje institucional y prevención de errores. En definitiva, su utilización mejora la eficiencia del sistema judicial al proporcionar un marco de referencia para la argumentación y la redacción, reduciendo los tiempos de análisis en casos similares. Finalmente, resulta relevante que la jurisprudencia se mantenga en constante actualización y se adapte a nuevas realidades sociales, contribuyendo a legitimar el sistema judicial, promover la transparencia y facilitar la comprensión de las decisiones judiciales.

2.1.2. Inteligencia artificial y modelo de lenguaje de gran tamaño

La inteligencia artificial (IA) es una tecnología que permite que las computadoras simulen la inteligencia y las capacidades humanas de resolución de problemas [3]. Su desarrollo ha experimentado un crecimiento exponencial, impulsado por factores como el aumento del poder computacional, la disponibilidad masiva de datos y la mejora en los algoritmos de aprendizaje automático, lo que ha permitido su incorporación en diversos ámbitos de la sociedad moderna, tales como la medicina, la educación, las finanzas, la ciencia y el sistema de justicia.
En este contexto, la aparición de grandes modelos de lenguaje, como GPT-4, Claude o LLaMA, ha contribuido a democratizar el uso de la inteligencia artificial, permitiendo que los usuarios interactúen con modelos capaces de comprender y generar lenguaje natural, asistir en tareas de programación, análisis de datos y búsquedas avanzadas. No obstante, es importante considerar que el fenómeno de la inteligencia artificial no puede reducirse a una cuestión puramente tecnológica, ya que sus aplicaciones influyen de forma creciente en los sistemas democráticos, las políticas públicas, la economía y los espacios sociales [4].
Los modelos de lenguaje de gran tamaño (Large Language Models, LLM) son entrenados con grandes volúmenes de datos para comprender, generar y completar texto de manera coherente [5]. En el ámbito jurídico, estos modelos permiten interpretar consultas formuladas en lenguaje natural y generar respuestas contextualmente relevantes basadas en información especializada. En este proyecto, se utiliza un LLM de código abierto ajustado para seguir instrucciones, integrado dentro de una arquitectura orientada a la generación de respuestas fundamentadas en evidencia documental, lo que resulta especialmente pertinente para el análisis de jurisprudencia laboral.

2.1.3. Generación aumentada por recuperación y búsqueda semántica

La generación aumentada por recuperación (Retrieval-Augmented Generation, RAG) es un proceso orientado a optimizar la salida de un modelo de lenguaje de gran tamaño, permitiendo que haga referencia a una base de conocimientos externa antes de generar una respuesta [6]. Esta estrategia resulta especialmente útil en escenarios donde el conocimiento se actualiza de forma constante, como ocurre en el ámbito jurídico, donde la jurisprudencia y las normativas pueden variar con frecuencia, contribuyendo a reducir el riesgo de respuestas no fundamentadas y asegurando que el modelo responda en base a documentos existentes.
La búsqueda semántica es una técnica orientada a comprender el significado de palabras y expresiones considerando la intención subyacente de una consulta realizada por un usuario [7]. Aplicada al contexto de la búsqueda de jurisprudencia, esta técnica permite obtener resultados más precisos y contextualizados, al

comprender conceptos jurídicos y sus variantes, aun cuando se utilicen sinónimos o expresiones distintas al lenguaje legal tradicional.
Para la implementación de este tipo de arquitecturas, existen herramientas que facilitan la integración de los distintos componentes del sistema. LangChain [8] es un framework que permite coordinar modelos de lenguaje, agentes [9] con roles definidos y estrategias de generación aumentada por recuperación, además de gestionar la interacción con bases de datos vectoriales. Por su parte, ChromaDB es una base de datos vectorial optimizada para la recuperación eficiente de información semántica, orientada a la gestión de embeddings y a la mejora de la precisión en la búsqueda de documentos relevantes. Gracias a estas características, ChromaDB [10] resulta especialmente adecuada para sistemas de análisis jurídico y búsqueda semántica de jurisprudencia, permitiendo recuperar fragmentos relevantes en función de su significado, uso y contexto.

2.2. Estado del Arte

En los últimos años, la aplicación de técnicas de inteligencia artificial al ámbito jurídico se ha consolidado como un área de creciente interés, con múltiples investigaciones orientadas a la predicción de decisiones judiciales, la búsqueda semántica y la recuperación automatizada de jurisprudencia. Estos trabajos evidencian avances relevantes en la automatización del análisis jurídico; sin embargo, también ponen de manifiesto limitaciones persistentes relacionadas con la escalabilidad de las soluciones, los costos asociados, la dependencia de servicios de terceros, la reserva de la información y la necesidad de validación experta. En este contexto, la revisión del estado del arte permite contextualizar la presente investigación y fundamentar la propuesta de solución desarrollada.

2.2.1. Búsqueda semántica de jurisprudencia basada en aprendizaje supervisado y semi-supervisado

Uno de los enfoques tempranos en la búsqueda semántica de jurisprudencia es el propuesto en A Semi-Supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases [11], donde se utiliza un método semi-supervisado para identificar hechos legales relevantes en sentencias de inmigración canadiense. En este trabajo se emplean word embeddings entrenados sobre un corpus especializado, combinados con un clasificador de oraciones que permite detectar hechos relevantes y casos similares mediante el uso de similitud coseno.

El corpus utilizado se compone de 150 documentos legales, los cuales fueron anotados manualmente por dos estudiantes de derecho, alcanzando un 90 % de precisión. No obstante, el enfoque presenta limitaciones significativas en términos de escalabilidad y costos, debido al trabajo manual requerido para la anotación, además de estar restringido a un dominio jurídico muy acotado. Asimismo, el uso de word embeddings junto a un clasificador con una sola capa oculta limita la capacidad del modelo para capturar relaciones semánticas profundas y contextos complejos presentes en las sentencias.

2.2.2. Comparación de modelos de embeddings y técnicas de fragmentación

Un enfoque más reciente es presentado en From Fact Draft to Operational Systems: Semantic Search in Legal Decisions Using Fact Drafts [12], donde se realiza una comparativa exhaustiva entre distintos modelos de embeddings aplicados a la búsqueda semántica de jurisprudencia húngara. En este estudio se analizan doce modelos, entre ellos Cohere, BGE-m3, OpenAI embeddings y Jina embeddings.

El corpus está compuesto por 1.172 sentencias judiciales, complementadas con resúmenes de hechos generados mediante un LLM y posteriormente corregidos de forma manual. El trabajo introduce técnicas relevantes para el manejo de sentencias extensas, tales como chunking, striding y Last Chunk Scaling (LCS), orientadas a mejorar la calidad de los vectores generados y la recuperación de información. Si bien el enfoque logra altos niveles de precisión, presenta limitaciones importantes, ya que depende de APIs comerciales como OpenAI o Cohere, lo que incrementa sustancialmente los costos y expone información sensible protegida por ley. Además, la evaluación se centra en métricas como MRR y Recall@k, orientadas principalmente a recuperar la sentencia original a partir de un borrador, sin reflejar completamente el desempeño ante consultas más diversas y complejas.

2.2.3. Modelado de temas y resumen automático de sentencias

Otro enfoque relevante es el presentado en Semantic Search and Summarization of Judgments Using Topic Modeling [13], donde se propone el uso de Latent Dirichlet Allocation (LDA) para identificar texto relevante en sentencias relacionadas con compensaciones por lesiones personales. Este enfoque incorpora, además, un resumen generado automáticamente que destaca los párrafos más pertinentes en función de la consulta del usuario.

El corpus utilizado está compuesto por 832 sentencias, categorizadas en antecedentes, lesiones, tratamientos y compensaciones. Para ello, se aplican tres variantes de modelado de temas: sin conocimiento de dominio, con conocimiento de características y con conocimiento de aspectos. No obstante, el uso de LDA presenta desventajas frente a modelos basados en embeddings más modernos, ya que no captura relaciones semánticas complejas ni dependencias profundas entre conceptos. Asimismo, este enfoque implica costos elevados y dificultades de escalabilidad, al requerir etiquetado manual por parte de expertos.

2.2.4. Arquitecturas RAG y predicción de decisiones judiciales

En NyayaRAG: Realistic Legal Judgment Prediction With RAG under the Indian Common Law System [14], se propone una arquitectura basada en RAG y LLM para la predicción de resultados judiciales en el sistema legal indio. El estudio utiliza un corpus de 50.000 casos y 5.000 resúmenes, evaluando distintos pipelines de recuperación, como la combinación de casos con leyes o normativa, y casos con jurisprudencia, los cuales son utilizados como contexto para modelos LLaMA-2 y variantes con fine-tuning.

Este trabajo busca no solo predecir el resultado de un caso, sino también ofrecer explicaciones sobre las razones de la decisión. Entre sus principales conclusiones, se destaca que la calidad del contexto recuperado influye directamente en las predicciones generadas. Además, se enfatiza que métricas automáticas como accuracy y F1-score no resultan suficientes por sí solas, subrayando la importancia de la validación por parte de expertos jurídicos para evaluar la coherencia y utilidad práctica de las respuestas.

2.2.5. Enfoques híbridos y conocimiento legal estructurado

El trabajo Bridging Legal Knowledge and AI: Retrieval-Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical Non-negative Matrix Factorization [15] propone un modelo híbrido para la recuperación de jurisprudencia, integrando arquitecturas RAG con conocimiento legal estructurado, como grafos normativos y ontologías jurídicas. El objetivo de este enfoque es entregar al modelo de lenguaje un contexto enriquecido, reduciendo errores normativos y mejorando la calidad de las respuestas generadas.

Si bien se trata de una propuesta metodológicamente interesante, su implementación requiere un esfuerzo humano considerable para la creación y mantención del conocimiento estructurado, capacidad que no se encuentra disponible en el contexto de esta tesina. No obstante, este tipo de aproximaciones se considera como una línea de trabajo futuro. En cuanto a la evaluación, los autores utilizan métricas como accuracy, F1-score y Recall, destacando nuevamente la relevancia de la validación humana para evaluar la exactitud y utilidad práctica de las respuestas.

2.2.6. Aplicaciones de IA en el ámbito judicial chileno y privado

En el contexto del Poder Judicial de Chile, la adopción de tecnologías basadas en inteligencia artificial ha sido progresiva y condicionada por limitaciones técnicas y presupuestarias. Entre las iniciativas desarrolladas se encuentran la anonimización automática de sentencias, la detección de órdenes de no innovar y, recientemente, la implementación de pilotos para la transcripción automática de audios de juicios de primera instancia, lo que contribuye a reducir los tiempos de redacción de los hechos.

En cuanto a la búsqueda de jurisprudencia, el Poder Judicial dispone de un portal público que permite realizar búsquedas mediante palabras clave. Si bien esta herramienta resulta útil para usuarios externos, presenta limitaciones relevantes para apoyar la redacción de sentencias, ya que no considera el contexto semántico ni identifica situaciones similares expresadas en términos distintos, generando resultados extensos que requieren revisión manual. En julio de 2025, se incorporó una mejora al buscador jurisprudencial mediante el uso de inteligencia artificial y consultas en lenguaje natural; sin embargo, durante las pruebas realizadas se observó que el sistema continúa basándose principalmente en coincidencias literales, sin un análisis semántico profundo, lo que limita su utilidad ante consultas complejas.

En el ámbito privado, existen diversas soluciones basadas en IA para aplicaciones jurídicas, entre las que destaca Sof-IA, desarrollada por la empresa española Tirant [16], orientada a apoyar a abogados en búsquedas normativas y jurisprudenciales, así como en la generación de textos y mapas conceptuales. No obstante, estas soluciones suelen depender de servicios externos y APIs de terceros, lo que plantea desafíos en términos de costos, confidencialidad y control de la información.

En comparación con los trabajos analizados, la propuesta desarrollada en esta investigación busca eliminar la dependencia de servicios externos, minimizar el trabajo manual y garantizar el manejo seguro de la información, combinando generación aumentada por recuperación con validación jurídica mediante prompts especializados y un enfoque centrado en el entendimiento semántico profundo del contexto jurídico.

3. Desarrollo

3.1. Especificación técnica de la solución

La solución propuesta se concibe para su despliegue en un entorno institucional controlado, utilizando infraestructura local de la institución, con el objetivo de resguardar la confidencialidad de la información y cumplir con los estándares de seguridad requeridos para el tratamiento de jurisprudencia. En este sentido, la arquitectura objetivo está diseñada para operar de manera autónoma, sin dependencia de servicios externos en un escenario productivo. No obstante, para las etapas de desarrollo, experimentación y validación, la solución fue implementada y evaluada mediante un Producto Mínimo Viable (MVP), utilizando el entorno de ejecución Google Colab, seleccionado exclusivamente por la disponibilidad de hardware y la facilidad de escalabilidad que ofrece para la ejecución de pruebas exploratorias y comparativas, sin que ello altere la concepción del despliegue final sobre infraestructura institucional.

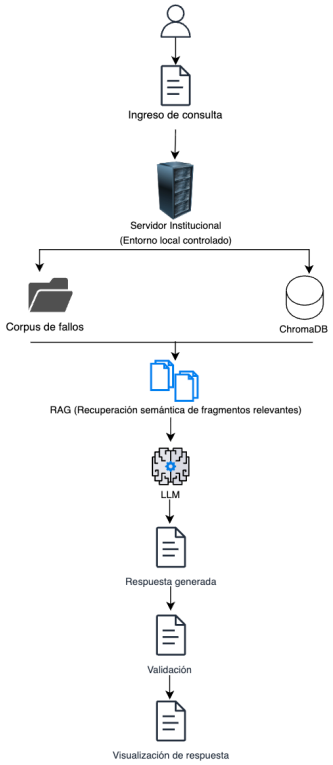


Figura 2. Arquitectura funcional del sistema propuesto de búsqueda de jurisprudencia

La Figura 2 presenta la arquitectura funcional del sistema propuesto de búsqueda de jurisprudencia, ilustrando la interacción entre los principales componentes del sistema —usuario, servidor institucional, corpus de fallos, base de datos vectorial ChromaDB, módulo RAG y modelo de lenguaje— así como su articulación general dentro de un entorno institucional local y controlado.

El corpus utilizado para la solución estuvo compuesto por 452 fallos en materia laboral emitidos por la Corte de Apelaciones de Iquique entre los años 2022 y 2025, los cuales fueron extraídos en formato PDF directamente desde el sistema de tramitación SITCORTE. La selección de este período respondió a criterios de estandarización del formato documental y a la necesidad de trabajar con jurisprudencia reciente y representativa. Previo a su incorporación al sistema, las sentencias fueron sometidas a un proceso de limpieza orientado a eliminar elementos no sustantivos, tales como logotipos, numeración de páginas, firmas digitales y otros artefactos que pudiesen introducir ruido en el proceso de representación vectorial.
Una vez depurados, los documentos fueron transformados a un formato estructurado (JSON), incorporando metadatos jurídicos relevantes para su trazabilidad y posterior análisis, tales como rol de la causa, año y archivo de origen. Posteriormente, cada sentencia fue segmentada mediante un proceso de *chunking*, aprovechando la estructura típica de las sentencias chilenas —"Vistos", "Considerando" y "Resuelve"— con el objetivo de preservar la coherencia jurídica de los fragmentos. Sobre estos fragmentos se generaron representaciones vectoriales utilizando el modelo de embeddings BAAI/bge-m3, seleccionado por su rendimiento en tareas de recuperación semántica en español, las cuales fueron almacenadas en una base de datos vectorial implementada con ChromaDB.

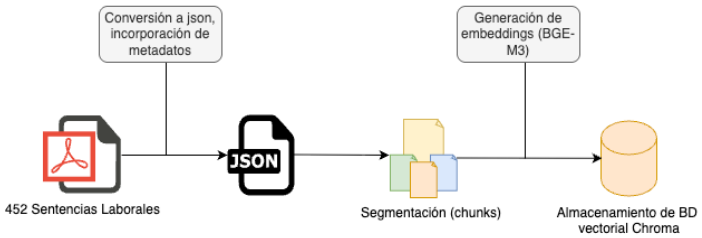


Figura 3. Flujo del tratamiento conversión y almacenamiento de los datos

La Figura 3 expone el flujo general de tratamiento y representación de los datos utilizado por el sistema, desde la ingestión de las sentencias en formato PDF hasta su estructuración, segmentación, vectorización y almacenamiento en la base de datos vectorial.

Sobre esta base de datos vectorial se diseñó e implementó el sistema utilizando una arquitectura de Generación Aumentada por Recuperación (RAG). El flujo funcional se inicia con la consulta del usuario en lenguaje natural, la cual es transformada en una representación vectorial mediante el modelo de embeddings BAAI/bge-m3. A continuación, dicho vector es consultado en ChromaDB para recuperar los fragmentos de jurisprudencia más pertinentes, junto con sus metadatos asociados, los cuales son incorporados como contexto en el *prompt* del modelo de lenguaje Qwen/Qwen2.5-14B-Instruct para la generación de una respuesta contextualizada y fundamentada en documentos del corpus.

Adicionalmente, el sistema incorpora un flujo de validación orientado a verificar que la respuesta generada se encuentre efectivamente respaldada por los fragmentos recuperados, mediante el uso de *prompts* especializados para las etapas de generación y verificación.

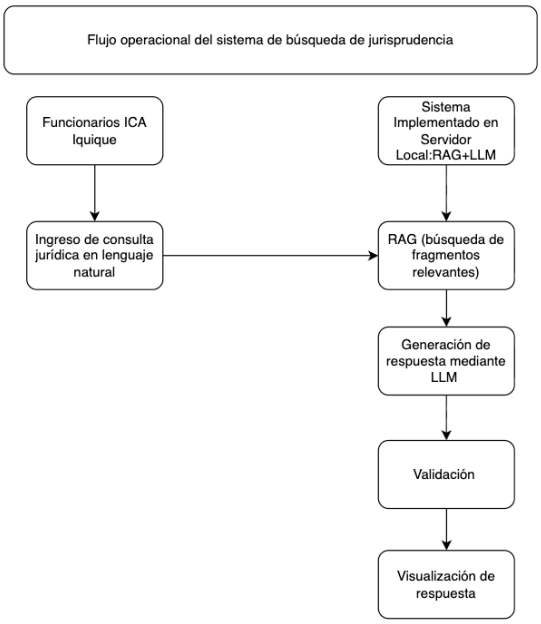


Figura 4. Flujo operacional del sistema propuesto

La Figura 4 describe el flujo operacional del sistema propuesto, detallando la secuencia de ejecución desde el ingreso de la consulta jurídica en lenguaje natural por parte del usuario, la recuperación semántica de fragmentos relevantes mediante la arquitectura RAG, la generación de la respuesta por el modelo de lenguaje, el proceso de validación y, finalmente, la visualización de la respuesta.

Finalmente, la justificación técnica y comparativa de la selección de los modelos empleados, así como las pruebas preliminares realizadas durante el desarrollo, se presentan en la siguiente sección.

3.2. Metodología y resultados

Una vez definida la especificación técnica de la solución propuesta, esta sección describe el proceso de preparación del corpus y las condiciones experimentales bajo las cuales se evaluó el sistema, incluyendo las etapas de preprocesamiento de los datos, experimentación preliminar y análisis de resultados.

3.2.1. Configuración final y procesamiento del corpus

El corpus utilizado para la evaluación de la propuesta estuvo compuesto por 452 fallos en materia laboral emitidos por la Corte de Apelaciones de Iquique entre los años 2022 y 2025, los cuales fueron extraídos en formato PDF directamente desde el sistema de tramitación SITCORTE. La selección de este período respondió a criterios de estandarización del formato documental y a la necesidad de trabajar con jurisprudencia reciente y representativa, alineada con las consultas habitualmente realizadas por los funcionarios del tribunal.
Previo a su incorporación al sistema, las sentencias fueron sometidas a un proceso de limpieza orientado a eliminar elementos no sustantivos, tales como logotipos, numeración de páginas, firmas digitales y otros artefactos que pudiesen introducir ruido en el proceso de representación vectorial. Una vez depurados, los documentos fueron transformados a un formato estructurado (JSON), incorporando metadatos jurídicos relevantes para su trazabilidad y posterior análisis, tales como rol de la causa, año y archivo de origen.
Posteriormente, cada sentencia fue segmentada mediante un proceso de chunking, aprovechando la estructura típica de las sentencias chilenas —"Vistos", "Considerando" y "Resuelve"— con el objetivo de preservar la coherencia jurídica de los fragmentos y evitar la mezcla de secciones con funciones argumentativas distintas. Sobre estos fragmentos se generaron representaciones vectoriales utilizando el modelo de embeddings BAAI/bge-m3, seleccionado por su desempeño en tareas de recuperación semántica en español. Dichos embeddings fueron almacenados en una base de datos vectorial implementada con ChromaDB, permitiendo la recuperación eficiente de fragmentos relevantes a partir de consultas en lenguaje natural.

3.2.2. Experimentos preliminares

El desarrollo del MVP se llevó a cabo mediante un proceso iterativo de experimentación, cuyo objetivo fue mejorar progresivamente la calidad de las respuestas generadas por el sistema y su utilidad práctica en el contexto de la búsqueda de jurisprudencia en la Corte de Apelaciones de Iquique. Durante estas etapas se realizaron pruebas preliminares sobre distintas configuraciones del sistema, considerando principalmente variaciones en el modelo de lenguaje, el modelo de embeddings y la estrategia de segmentación de documentos.

En una etapa inicial, el MVP fue evaluado utilizando un corpus preliminar compuesto por 933 fallos laborales emitidos por la Corte de Apelaciones de Iquique durante el período 2016–2022, con el propósito de establecer una referencia inicial de desempeño. Sin embargo, las pruebas evidenciaron que dicho corpus no se encontraba plenamente alineado con las consultas jurisprudenciales formuladas por los usuarios expertos, orientadas mayoritariamente a criterios y prácticas más recientes. En función de lo anterior, se optó por trabajar con un corpus actualizado de 452 fallos laborales correspondientes al período 2022–2025.

Asimismo, durante las fases iniciales de experimentación se realizaron pruebas exploratorias con modelos de lenguaje cuantizados, las cuales fueron descartadas tempranamente debido a su bajo desempeño en términos de coherencia y exactitud jurídica.

Posteriormente, las pruebas preliminares se concentraron en tres modelos de lenguaje de código abierto: Mistral-7B-Instruct, Meta-Llama-3.1-8B-Instruct y Qwen2.5-14B-Instruct. Estas evaluaciones se realizaron dentro del mismo pipeline RAG y sobre el mismo corpus jurídico, lo que permitió comparar de manera directa el desempeño de los modelos en condiciones equivalentes. La selección progresiva del modelo generador se

sustentó exclusivamente en los resultados empíricos obtenidos durante estas pruebas, priorizando métricas automáticas y validación experta aplicadas al propio MVP.

3.2.3. Resultados del sistema propuesto

La evaluación final del sistema se realizó mediante un enfoque híbrido que combinó métricas automáticas reference-free y validación experta, permitiendo analizar el desempeño del MVP tanto desde una perspectiva cuantitativa como cualitativa, así como evaluar su comportamiento sobre distintos corpus. Este enfoque permitió obtener evidencia complementaria respecto de la coherencia textual, la calidad semántica de las respuestas y su utilidad práctica en el contexto jurídico.

En primer lugar, se utilizó la métrica BLANC [17] (Ver Anexo N° 4) para evaluar la coherencia entre los fragmentos de jurisprudencia recuperados y las respuestas generadas por el sistema, midiendo el aporte efectivo del modelo generador al contexto recuperado sin requerir anotaciones manuales. Para cada consulta se consideraron diez fragmentos de jurisprudencia, calculándose posteriormente un valor promedio por pregunta. Los resultados obtenidos muestran valores de BLANC estables, lo que indica que el modelo generador mantiene un comportamiento consistente incluso al ser aplicado sobre datos no utilizados durante el proceso de ajuste del sistema.

En segundo lugar, se aplicó la métrica G-EVAL mediante un comité compuesto por modelos de lenguaje de distintas familias , DeepSeek-7B, Gemma-2-9B y Yi-9B (Ver Anexo N° 3), con el objetivo de evaluar la calidad de las respuestas generadas por los modelos Qwen2.5-14B-Instruct, Meta-Llama-3.1-8B-Instruct y Mistral-7B-Instruct v0.2 en términos de pertinencia, exactitud normativa y contextual, valor para fundamentar una decisión y claridad.
La utilización de evaluadores externos a las familias de los modelos generadores permitió reducir posibles sesgos estructurales en la evaluación. Los resultados obtenidos reflejan un comportamiento consistente del sistema, observándose promedios similares entre los distintos escenarios evaluados, lo que refuerza la estabilidad del MVP frente a variaciones en el corpus.

Adicionalmente, se realizó una validación experta con funcionarios de la unidad de secretaría de ministros de la Corte de Apelaciones de Iquique (Ver Anexo N°2), quienes evaluaron las respuestas generadas por el sistema en un contexto de uso real (Ver Anexo N° 8). Esta validación se llevó a cabo mediante la aplicación de un cuestionario tipo Likert [19] (Ver Anexo N° 7), diseñado específicamente para este estudio y aplicado a un conjunto de 19 consultas jurídicas en materia laboral, formuladas por los propios usuarios expertos y alineadas con las consultas habitualmente realizadas durante la búsqueda de jurisprudencia. Cada respuesta fue evaluada considerando los criterios de pertinencia jurisprudencial, exactitud normativa y contextual, valor para fundamentar una decisión y claridad y usabilidad, utilizando una escala de valoración de 1 a 5. Los resultados muestran un alto nivel de utilidad práctica del MVP, confirmando su capacidad para apoyar efectivamente la búsqueda de jurisprudencia en escenarios institucionales reales.

Con el objetivo de facilitar la interpretación global de los resultados, la Tabla 2 presenta un resumen consolidado de los principales indicadores obtenidos para los tres modelos de lenguaje evaluados (Mistral-7B-Instruct v0.2, Meta-Llama-3.1-8B-Instruct y Qwen2.5-14B-Instruct), integrando los valores promedio de BLANC, G-EVAL y validación experta

Tabla 2. Resultados consolidados del sistema

Modelo	Blanc	G-Eval(Comité LLM's)	Comité de Expertos
Qwen 2.5 14B Instruct	-0.0316	3.850	4.675
Meta Llama 3.1 8B Instruct	-0.0868	3.570	2.298
Mistral 7B Instruct v0.2	-0.2318	3.706	1.728

Nota: Los valores corresponden a promedios obtenidos a partir de 19 consultas jurídicas en materia laboral. Todos los experimentos fueron realizados utilizando el modelo de embeddings BAAI/bge-m3 y la misma configuración de recuperación.

Acorde a los resultados, el modelo Mistral-7B-Instruct v0.2 presentó un desempeño inferior en la métrica BLANC y en la validación experta, a pesar de mostrar un rendimiento competitivo en la métrica G-EVAL, donde superó al modelo Meta-Llama-3.1-8B-Instruct. No obstante, estos resultados no fueron suficientes para compensar sus limitaciones en coherencia y utilidad práctica, lo que refuerza su descarte como modelo generador final.

Por su parte, Meta-Llama-3.1-8B-Instruct mostró resultados intermedios y un comportamiento estable en las distintas métricas evaluadas, aunque consistentemente por debajo del modelo finalmente seleccionado.

En conjunto, estos resultados respaldan la selección de Qwen/Qwen2.5-14B-Instruct como modelo generador final del MVP, decisión basada en evidencia empírica obtenida mediante métricas automáticas y validación experta.

3.2.3.1. Nivel de acuerdo inter-juez

Con el fin de analizar la consistencia en la evaluación de las respuestas generadas por el sistema, se aplicó la métrica de acuerdo inter-juez Kappa de Fleiss [20] tanto al comité de modelos de lenguaje como al comité de expertos humanos. Esta diferenciación resulta relevante, dado que ambos comités responden a lógicas evaluativas distintas: mientras los modelos de lenguaje presentan variabilidad asociada a diferencias en arquitectura, entrenamiento y sesgos implícitos, los expertos humanos evalúan desde una perspectiva jurídico-práctica, donde la interpretación normativa admite matices.

En el caso del comité de LLMs, los valores de Kappa obtenidos se situaron en rangos de acuerdo leve a aceptable, lo que resulta coherente con la heterogeneidad de los modelos evaluadores utilizados. Destaca que el modelo Qwen/Qwen2.5-14B-Instruct alcanzó el mayor nivel de acuerdo dentro de este comité, lo que sugiere una mayor consistencia en la evaluación de sus respuestas incluso frente a criterios automáticos diversos.

Tabla 3. Resultados de acuerdo inter-juez – Comité de LLMs

Modelo Evaluado	Kappa de Fleiss - comité de LLMs	Nivel de Acuerdo
Qwen 2.5 14B Instruct	0.3202	Acuerdo aceptable
Meta Llama 3.1 8B Instruct	0.1283	Acuerdo leve
Mistral 7B Instruct v0.2	0.2544	Acuerdo Aceptable

Por su parte, el comité de 3 funcionarios expertos presentó niveles de acuerdo moderado para los tres modelos evaluados. Este resultado es consistente con el tipo de tarea analizada, ya que la evaluación de jurisprudencia implica razonamiento jurídico, ponderación de criterios y experiencia profesional, factores que naturalmente generan variabilidad interpretativa. En este contexto, los valores obtenidos refuerzan la confiabilidad de la validación experta y respaldan la solidez metodológica del proceso evaluativo aplicado al MVP.

Tabla 4. Resultados de acuerdo inter-juez - Comité de Expertos

Modelo Evaluado	Kappa de Fleiss - comité de expertos	Nivel de Acuerdo
Qwen 2.5 14B Instruct	0.4478	Acuerdo moderado
Meta Llama 3.1 8B Instruct	0.5669	Acuerdo Moderado
Mistral 7B Instruct v0.2	0.5655	Acuerdo Moderado

3.2.4. Prueba de transferibilidad del sistema

Con el objetivo de evaluar la capacidad de transferencia del sistema propuesto, el MVP fue aplicado sobre un corpus independiente de 452 sentencias laborales correspondientes a la Corte de Apelaciones de Arica, el cual no fue utilizado durante ninguna etapa del desarrollo ni ajuste del sistema.

La Tabla 5 presenta una comparación de los resultados obtenidos para el corpus base de Iquique y el corpus de Arica, considerando las métricas BLANC, G-EVAL y validación experta.

Tabla 5. Resultados de la prueba de transferibilidad del sistema

Corpus	Blanc	G-Eval(Comité LLM's)	Comité de Expertos
Iquique	-0.0316	3.850	4.675
Arica	-0.0936	3.846	4.486

Nota: La prueba de transferibilidad fue aplicada exclusivamente sobre la configuración final del sistema, utilizando el modelo generador Qwen/Qwen2.5-14B-Instruct, el modelo de embeddings BAAI/bge-m3 y la misma estrategia de recuperación y evaluación empleada en el corpus base de Iquique.

Los resultados evidencian un comportamiento consistente del sistema al ser aplicado sobre un corpus no utilizado durante su desarrollo. En términos de coherencia textual, medida mediante la métrica BLANC, se observa una disminución moderada respecto del corpus base, sin que ello afecte de manera significativa la calidad global del sistema. Asimismo, las métricas G-EVAL y validación experta presentan valores prácticamente equivalentes entre ambos corpus, lo que indica que el MVP mantiene su calidad semántica y utilidad práctica frente a datos no vistos.

Adicionalmente, se evaluó el nivel de acuerdo inter-juez sobre el corpus de Arica mediante la métrica Kappa de Fleiss, considerando tanto el comité de modelos de lenguaje como el comité de expertos humanos. La Tabla 6 presenta los valores obtenidos para cada comité evaluador.

Tabla 6. Resultados de acuerdo inter-juez – Corpus Arica

Evaluación	Kappa de Fleiss - comité de LLMs	Kappa de Fleiss - comité de expertos
Arica	0.2654	0.5204

Los resultados muestran un valor de Kappa de 0.2654 para el comité de LLMs, correspondiente a un acuerdo aceptable, y un valor de 0,5204 para el comité de expertos, ubicado en el rango de acuerdo moderado. Estos niveles de concordancia son consistentes con los observados en el corpus base de Iquique, lo que indica que la variabilidad entre evaluadores se mantiene estable al aplicar el sistema sobre un conjunto de datos no utilizado durante su desarrollo.

La mantención de niveles de acuerdo similares en ambos corpus refuerza la confiabilidad del proceso evaluativo y respalda la transferibilidad del MVP, evidenciando que el sistema no solo conserva su desempeño técnico, sino también la coherencia en la evaluación de sus respuestas en contextos institucionales distintos.

En conjunto, estos resultados confirman la capacidad de generalización del sistema propuesto y respaldan su transferibilidad a contextos institucionales distintos, sin degradaciones relevantes en coherencia, calidad ni utilidad jurídica.

4. Conclusiones

La implementación de un sistema de inteligencia artificial que basado modelos de lenguaje y técnicas de generación aumentada de recuperación (RAG), demostró al desarrollar esta investigación que es una solución viable para optimizar el actual método de trabajo utilizado en la Corte de Apelaciones de Iquique. Mediante el MVP se evidenció que estas tecnologías pueden apoyar eficientemente a los funcionarios en una tarea que requiere análisis y precisión jurisprudencial.

Los resultados obtenidos mediante métricas BLANC y G-EVAL, además de la evaluación empírica realizada por los usuarios expertos demostraron una mejora sustancial en la calidad de las respuestas y una reducción significativa en el esfuerzo y tiempo necesario para buscar jurisprudencia relevante. Los participantes destacaron positivamente la experiencia y valoraron la utilidad del sistema al entregar un resumen de los fallos y los roles específicos.

Teniendo como base los resultados obtenidos y lo expresado por los usuarios expertos, se confirma integralmente la hipótesis planteada en esta investigación. El sistema propuesto generó respuestas de calidad suficiente para su uso en contextos judiciales reales, evidenciado por puntajes satisfactorios en métricas automáticas no supervisadas —con un promedio G-EVAL de 3,85 sobre 5— y una valoración positiva de los expertos jurídicos mediante escala Likert —con un promedio de 4,7 sobre 5—. Adicionalmente, se confirmó una reducción significativa en los tiempos de búsqueda: los usuarios expertos indicaron tiempos de búsqueda manual en el rango de 60 a 120 minutos, mientras que con el sistema propuesto dicho tiempo se redujo a un rango aproximado de 30 a 60 minutos por consulta, lo que representa una disminución cercana al 50 %, superando lo planteado inicialmente.

Al comparar los resultados de esta investigación con el estado del arte, particularmente con NyayaRAG, se observa consistencia en el orden de magnitud y comportamiento de la métrica BLANC, la cual presenta valores cercanos a cero, tal como se reporta en trabajos previos. Cabe destacar que BLANC no se interpreta como una métrica de desempeño absoluto, sino como una medida del aporte contextual del modelo generador sobre los fragmentos recuperados. En este sentido, el modelo seleccionado para la configuración final del MVP presenta el comportamiento más estable y con menor alteración del contexto recuperado, manteniendo valores de BLANC consistentes entre corpus. Estos resultados refuerzan la validez de la propuesta y evidencian un comportamiento coherente con lo reportado en la literatura. Asimismo, los hallazgos confirman la necesidad de complementar esta métrica con otros enfoques de evaluación, como G-EVAL, la validación experta y el análisis de acuerdo inter-juez, los cuales permiten capturar dimensiones semánticas, jurídicas y de utilidad práctica que BLANC, por sí sola, no aborda.

Finalmente, los resultados obtenidos por el MVP invitan a seguir investigando en implementaciones con inteligencia artificial que apoyen y fortalezcan la labor judicial y promuevan la modernización del sistema judicial chileno.

4.1. Trabajo Futuro

4.1.1 Ampliación del corpus a otras materias: Si bien la implementación actual consideró solo fallos laborales, es importante continuar escalando a otras materias como civil, penal o familia donde el volumen de información es aún mayor que en la materia actual.

4.1.2 Entrenamiento Fino: Algo altamente interesante sería realizar ajuste fino del modelo, lo que podría mejorar aún mas la calidad y coherencia de las respuestas.

4.1.3 Incorporar conocimiento legal estructurado de normativa y leyes: Incorporar la validación normativa permitiría al sistema no solo realizar búsqueda sino también evaluar los fundamentos jurídicos, disminuyendo los errores, detectando afirmaciones no validas en los documentos recuperados.

4.1.4 Interfaz de usuario y despliegue: Finalmente, desarrollar una interfaz completa que otorgue acceso a más usuarios al sistema, en conjunto a la implementación en servidores de nivel central.

5. Referencias

[1] M. R. Herrera Carbuccia, "La Sentencia", Gaceta Laboral, vol. 14, no. 1, pp. 133-156, 2008, [En línea]. Disponible en: http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1315-85972008000100006&lng=es&tlng=es

[2] Biblioteca del Congreso Nacional de Chile, "Glosario jurídico", 2024, [En línea]. Disponible en: https://www.bcn.cl/leychile/glosario

[3] IBM, "¿Qué es la inteligencia artificial (IA)?", 2024, [En línea]. Disponible en: https://www.ibm.com/mx-es/topics/artificial-intelligence

[4] Pontificia Universidad Católica de Chile, "¿Qué impactos socioculturales tienen los usos de la inteligencia artificial?", 2024, [En línea]. Disponible en: https://www.uc.cl/noticias/header-que-impactossocioculturales-tienen-los-usos-de-la-inteligencia-artificial/

[5] Amazon, "¿Qué son los modelos de lenguaje de gran tamaño?", 2025, [En línea]. Disponible en: https://aws.amazon.com/es/what-is/large-language-model/

[6] Amazon Web Services, "¿Qué es la generación aumentada por recuperación (RAG)?", 2024, [En línea]. Disponible en: https://aws.amazon.com/es/what-is/retrieval-augmented-generation/

[7] GoogleCloud, "¿Qué es la búsqueda semántica?", 2025, [En línea]. Disponible en: https://cloud.google.com/discover/what-is-semantic-search?hl=es

[8] LangChain, "LangChain Documentation", 2024, [En línea]. Disponible en: https://www.langchain.com/

[9] IBM, "What is an Intelligent Agent?", 2024, [En línea]. Disponible en: https://www.ibm.com/cloud/learn/intelligent-agent

[10] ChromaDB, "Chroma Documentation", 2024, [En línea]. Disponible en: https://docs.trychroma.com/docs/overview/introduction

[11] M. Nejadgholi, F. Gonzalez, G. Hirst, and C. van Kessel, "A Semi-Supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases," *Frontiers in Artificial Intelligence and Applications*, vol. 302, pp. 123-132, 2017.

[12] G. Csányi, D. Görcs, and A. Micsik, "From Fact Drafts to Operational Systems: Semantic Search in Legal Decisions Using Fact Drafts," *Big Data and Cognitive Computing*, vol. 8, no. 185, 2024.

[13] T.-H. Wu, B. Kao, F. Chan, A. S. Y. Cheung, M. M. K. Cheung, G. Yuan, and Y. Chen, "Semantic Search and Summarization of Judgments Using Topic Modeling," *Legal Knowledge and Information Systems*, vol. 346, pp. 100-106, 2021.

[14] A. Bhattacharya, S. Ghosh, and L. Dey, "NyayaRAG: Retrieval Augmented Generation for Indian Legal Documents," arXiv preprint arXiv:2309.05675, 2023. [En línea]. Disponible en: https://arxiv.org/abs/2309.05675

[15] R. C. Barron, M. E. Eren, O. M. Serafimova, C. Matuszek, and B. S. Alexandrov, "Bridging Legal Knowledge and AI: Retrieval-Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical

Non-negative Matrix Factorization," arXiv preprint arXiv:2502.20364, 2025. [En línea]. Disponible en: https://arxiv.org/abs/2502.20364

[16] Tirant Lo Blanch, "Sof-IA: La IA para abogados más eficiente," 2024, [En línea]. Disponible en:https://prime.tirant.com/cl/actualidad-prime/sof-ia-la-ia-para-abogados-mas-eficiente/

[17] O. Vasilyev, V. Dharnidharka y J. Bohannon, "BLANC: Human-free quality estimation of document summaries," arXiv preprint arXiv:2002.09836v2, 2020. [En línea]. Disponible: https://arxiv.org/abs/2002.09836

[18] Y. Liu, D. Iter, J. Wang, et al., "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," arXiv preprint arXiv:2303.16634, 2023. [En línea]. Disponible en: https://arxiv.org/abs/2303.16634

[19] Questionpro, "¿Qué es la escala likert y cómo utilizarla?" 2025, [En línea]. Disponible en: https://www.questionpro.com/blog/es/que-es-la-escala-de-likert-y-como-utilizarla/

[20] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159–174, 1977. doi: 10.2307/2529310. [En línea]. Disponible en: https://www.jstor.org/stable/2529310