UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA, INFORMÁTICA Y MECÁNICA

ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA Y DE SISTEMAS

TESIS

AUTOMATIZACIÓN DE PRE-DICTÁMENES JURÍDICOS CON IA EN EL ÁREA DE ASESORÍA LEGAL DE LA MUNICIPALIDAD DE CALCA

PRESENTADO POR:
Br. JHON EDWIN TACUSI LAROTA
Br. ALEXANDER TTITO SAYA

PARA OPTAR AL TÍTULO PROFESIONAL DE INGENIERO INFORMATICO Y DE SISTEMAS

ASESOR
Mgt. JOSE MAURO PILLCO QUISPE

CUSCO – PERU

2026

Universidad Nacional de San Antonio Abad del Cusco

INFORME DE SIMILITUD

(Aprobado por Resolución Nro.CU-321-2025-UNSAAC)

El que suscribe, el **Asesor** JOSÉ MAURO PILLCO QUISPE quien aplica el software de detección de similitud al trabajo de investigación/tesis titulada: AUTOMATIZACIÓN DE PCE - DICTÁMENES JURÍDICOS CON IA EN EL ÁREA DE ASESORIA LEGAL DE LA MUNICIPALIDAD DE CALCA

Presentado por: JHON EDWIN TACUSI LAROTA DNI N° 72212772;
presentado por: ALEXANDER TITO SAYA DNI N°: 75757579
Para optar el título Profesional/Grado Académico de INGENIERO INFORMÁTICO Y DE SISTEMAS

Informo que el trabajo de investigación ha sido sometido a revisión por 3 veces, mediante el Software de Similitud, conforme al Art. 6° del ***Reglamento para Uso del Sistema Detección de Similitud en la UNSAAC*** y de la evaluación de originalidad se tiene un porcentaje de 5 %.

Evaluación y acciones del reporte de coincidencia para trabajos de investigación conducentes a grado académico o título profesional, tesis

Porcentaje	Evaluación y Acciones	Marque con una (X)
Del 1 al 10%	No sobrepasa el porcentaje aceptado de similitud.	X
Del 11 al 30 %	Devolver al usuario para las subsanaciones.	
Mayor a 31%	El responsable de la revisión del documento emite un informe al inmediato jerárquico, conforme al reglamento, quien a su vez eleva el informe al Vicerrectorado de Investigación para que tome las acciones correspondientes; Sin perjuicio de las sanciones administrativas que correspondan de acuerdo a Ley.	

Por tanto, en mi condición de Asesor, firmo el presente informe en señal de conformidad y **adjunto** las primeras páginas del reporte del Sistema de Detección de Similitud.

Cusco, 09 de Octubre de 2026

Firma

Post firma JOSÉ MAURO PILLCO QUISPE

Nro. de DNI 23861067

ORCID del Asesor 0000-0002-0527-089X

Se adjunta:

1. Reporte generado por el Sistema Antiplagio.
2. Enlace del Reporte Generado por el Sistema de Detección de Similitud: **oid:** 27259:636694222

ALEXANDER TTITO JHON EDWIN TACUSI

Tesis alexander Ttito Jhon Tacusi.pdf

Universidad Nacional San Antonio Abad del Cusco

Detalles del documento

Identificador de la entrega
trn:oid:::27259:636694222

Fecha de entrega
9 oct 2026, 12:22 p.m. GMT-5

Fecha de descarga
9 oct 2026, 12:34 p.m. GMT-5

Nombre del archivo
Tesis alexander Ttito Jhon Tacusi.pdf

Tamaño del archivo
7.4 MB

190 páginas

38.109 palabras

237.640 caracteres

5% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado
- Coincidencias menores (menos de 15 palabras)

Exclusiones

- N.º de coincidencias excluidas

Fuentes principales

5% Fuentes de Internet
2% Publicaciones
4% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Dedicatoria

Dedico este trabajo, con gratitud, a mis padres y a mi hermano, quienes han sido un pilar fundamental a lo largo de este camino. Su amor, confianza y apoyo incondicional me brindaron la fortaleza necesaria para afrontar cada desafío y no desistir ante las dificultades. Cada palabra de aliento y cada muestra de confianza fueron una fuente de motivación para continuar avanzando y alcanzar esta meta, que hoy representa también un logro compartido con ellos.

Br. Jhon Edwin Tacusi Larota

Dedico este trabajo especialmente a mi familia (padres y hermanos). Ellos han sido mi principal fuente de motivación para alcanzar mis objetivos. Su apoyo incondicional de mi mamá ha sido fundamental a lo largo de esta etapa, especialmente en los momentos más difíciles, cuando sus palabras de aliento me dieron la fuerza necesaria para seguir adelante.

Br. Alexander Ttito Saya

Agradecimientos

A mi asesor de tesis, Mgt. José Mauro Pillco Quispe, por su orientación, paciencia y compromiso en cada etapa de esta tesis.

A los docentes de la Escuela Profesional de Ingeniería Informática y de Sistemas, quienes compartieron sus conocimientos y experiencia a lo largo de toda mi formación universitaria.

A la Escuela Profesional de Ingeniería Informática y de Sistemas, espacio donde adquirí los fundamentos y herramientas que guían mi ejercicio profesional.

A mi tricentenaria casa de estudios, la Universidad Nacional de San Antonio Abad del Cusco, institución donde se forjan el conocimiento, la vocación y el compromiso con nuestra sociedad.

Resumen

La Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca presenta limitaciones en la elaboración oportuna de opiniones legales debido al incremento de documentos administrativos y la insuficiencia de personal jurídico, situación asociada a restricciones normativas y presupuestales establecidas en la Ley N.º 31298 y el Memorándum Múltiple N.º 0277-2023-GM-MPC.

La investigación tuvo como objetivo determinar la eficacia de una solución basada en inteligencia artificial generativa, mediante modelos de lenguaje a gran escala (LLM), para automatizar la elaboración de opiniones legales. Se desarrolló una investigación aplicada, de nivel descriptivo-explicativo y diseño cuasi-experimental. Se compararon los modelos Claude, Llama y Mistral mediante métricas de similitud lexica y semántica, así como el tiempo de generación.

Los resultados mostraron que Claude obtuvo el mejor desempeño global, destacando en BERT Precision (0.739), BERT Recall (0.735), BERT F1 (0.737) y ROUGE-2 (0.201), además del menor tiempo promedio de generación (33.81 segundos), frente a Llama (105.96 segundos) y Mistral (114.09 segundos). Por ello, fue seleccionado para desarrollar un prototipo funcional que automatiza el procesamiento de expedientes administrativos y la generación de opiniones legales. Se emplearon técnicas de ingeniería de prompts, como Plan and Solve Prompting y Chain of Verification Prompting. Finalmente, seis abogados especialistas evaluaron el prototipo mediante una ficha con escala Likert de cinco niveles. Se concluye que los LLM constituyen una alternativa viable para apoyar la elaboración de opiniones legales, sin sustituir el criterio profesional.

Palabras Clave: *Inteligencia artificial generativa, Modelos de lenguaje a gran escala (LLM), Opiniones legales, Ingeniería de prompts, Automatización jurídica.*

Abstract

The Legal Advisory Office of the Calca Provincial Municipality faces limitations in the timely preparation of legal opinions due to an increase in administrative documents and a shortage of legal staff—a situation linked to regulatory and budgetary restrictions established by Law No. 31298 and Multiple Memorandum No. 0277-2023-GM-MPC.

The study aimed to determine the effectiveness of a solution based on generative artificial intelligence—specifically large language models (LLMs)—to automate the drafting of legal opinions. An applied research study was conducted using a descriptive-explanatory approach and a quasi-experimental design. The Claude, Llama, and Mistral models were compared based on lexical and semantic similarity metrics, as well as generation time.

The results showed that Claude achieved the best overall performance, excelling in BERT Precision (0.739), BERT Recall (0.735), BERT F1 (0.737), and ROUGE-2 (0.201), while also recording the shortest average generation time (33.81 seconds) compared to Llama (105.96 seconds) and Mistral (114.09 seconds). Consequently, it was selected to develop a functional prototype that automates the processing of administrative files and the generation of legal opinions. Prompt engineering techniques, such as "Plan and Solve Prompting" and "Chain of Verification Prompting," were employed. Finally, six specialized lawyers evaluated the prototype using a five-point Likert scale. The study concludes that LLMs represent a viable alternative to support the drafting of legal opinions without replacing professional judgment..

Keywords: *Generative artificial intelligence, Extended language models (LLM), Legal reports, Prompt engineering, Legal automation.*

Introducción

En la actualidad, a nivel mundial, la transformación digital ha generado cambios significativos en la forma en que las organizaciones desarrollan sus actividades, impulsando la incorporación de tecnologías innovadoras como la Inteligencia Artificial (IA). Estas herramientas han permitido optimizar procesos, reducir tiempos de trabajo y mejorar la calidad de los servicios en distintos sectores, tales como la salud, la educación, la industria y la administración pública. En particular, la Inteligencia Artificial Generativa, a través de modelos de lenguaje de gran escala (LLM), ha demostrado ser una solución eficiente para automatizar tareas que implican el procesamiento de grandes volúmenes de información y la generación de contenido especializado.

En el ámbito jurídico, estas tecnologías han comenzado a utilizarse para apoyar actividades como la revisión de documentos, la búsqueda de normativa, el análisis de antecedentes y la redacción de opiniones legales. Su aplicación permite agilizar el trabajo profesional, disminuir la carga operativa y mejorar la calidad de los documentos generados. Sin embargo, su implementación en el sector público aún es limitada, especialmente en países en desarrollo, debido a factores como restricciones normativas, limitaciones presupuestales y brechas tecnológicas que dificultan su adopción.

En el Perú, si bien existen avances en materia de gobierno digital, muchas entidades públicas aún enfrentan dificultades para modernizar sus procesos internos. Esto se traduce en procedimientos administrativos lentos, sobrecarga laboral del personal y limitaciones en la atención oportuna de las demandas ciudadanas. En el caso específico del sector jurídico dentro de las entidades públicas, estas limitaciones se hacen más evidentes debido a la complejidad de los documentos legales que deben analizarse, así como a la necesidad de emitir opiniones legales precisas, fundamentadas y dentro de plazos establecidos.

En este contexto, la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca no es ajena a esta problemática. Dicha área cumple un rol fundamental dentro de la institución, ya que es responsable de analizar, evaluar y emitir opiniones legales sobre diversos asuntos administrativos y normativos. No obstante, en los últimos años se ha evidenciado un incremento significativo en la carga de trabajo, debido al aumento de documentos que requieren revisión, tales como expedientes administrativos, informes técnicos, resoluciones y normativas aplicables.

A esta situación se suma la limitada disponibilidad de personal jurídico especializado, lo que dificulta atender de manera oportuna todos los requerimientos. Como consecuencia, se generan retrasos en la emisión de opiniones legales, así como posibles deficiencias en su elaboración. Esta problemática ha sido evidenciada en diversos documentos internos, como el Informe N.º 029-2023-MPC-GM/OAJ y el Informe N.º 135-2023-MPC-GM/OAJ, en los cuales se señala la necesidad de incorporar mayor personal especializado para mejorar el desempeño del área.

Sin embargo, la posibilidad de ampliar el equipo de trabajo se encuentra restringida por disposiciones normativas y presupuestales. Entre ellas destaca la Ley N.º 31298, que limita determinadas modalidades de contratación, así como el Memorándum Múltiple N.º 0277-2023-GM-MPC/JLVG, orientado a la racionalización del gasto institucional. Estas restricciones han impedido la contratación de nuevo personal, incrementando aún más la carga laboral existente y afectando la eficiencia del servicio jurídico.

Frente a este escenario, surge la necesidad de buscar alternativas innovadoras que permitan optimizar los procesos sin depender exclusivamente del aumento de recursos humanos. En este sentido, la Inteligencia Artificial Generativa se presenta como una herramienta viable para apoyar las funciones de la Oficina de Asesoría Jurídica,

especialmente en tareas repetitivas y de alto volumen documental. Estas incluyen la revisión de antecedentes, la identificación de normativa aplicable y la elaboración preliminar de opiniones legales.

El uso de modelos de lenguajes extensos puede contribuir a reducir los tiempos de atención, mejorar la organización de la información y optimizar la precisión en la redacción de documentos legales. No obstante, es importante señalar que estas herramientas no reemplazan la labor del profesional jurídico, sino que actúan como un apoyo que permite mejorar su productividad, siempre bajo supervisión humana y en cumplimiento del marco normativo vigente.

Diversas investigaciones, como las desarrolladas por Ashley y Phelps (2022) y Mongoli (2024), destacan que la aplicación de la inteligencia artificial en el ámbito jurídico permite no solo agilizar los procesos, sino también mejorar la calidad del trabajo profesional. Asimismo, coinciden en la importancia de garantizar la revisión humana de los resultados generados, a fin de asegurar su validez, coherencia y adecuación a las normas legales.

En atención a lo expuesto, el presente trabajo de investigación tiene como propósito diseñar e implementar una solución tecnológica basada en Inteligencia Artificial Generativa en la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca. Esta propuesta busca optimizar el proceso de elaboración de opiniones legales, reducir los tiempos de atención y mejorar la eficiencia y productividad del área, contribuyendo así a la modernización de la gestión pública y al fortalecimiento de los servicios brindados por la institución.

Índice

Capítulo I: Aspectos generales 1

1.1 Planteamiento de Problema 1

1.1.1 Descripción del problema 1

1.2 Formulación del problema 3

1.2.1 Problema general 3

1.2.2 Problemas específicos 3

1.3 Objetivos 3

1.3.1 Objetivo general 3

1.3.2 Objetivos específicos 4

1.4 Justificación 4

1.4.1 Justificación tecnológica 4

1.4.2 Justificación Teórica 5

1.4.3 Justificación Práctica 6

1.4.4 Justificación Metodológica 6

1.4.5 Justificación Social 7

1.4.6 Justificación Legal 7

1.4.7 Justificación Económica 8

1.5 Alcances y limitaciones 9

1.5.1 Alcances 9

1.5.2 Limitaciones 9

Capítulo II: Marco teórico 11

2.1 Antecedentes de la Investigación 11

2.1.1 Antecedentes Internacionales 11

2.1.2 Antecedentes Nacionales 17

2.2 Bases Teóricas 19

2.2.1 Inteligencia artificial (IA) 19

2.2.2 Inteligencia Artificial en la Administración Pública 20

2.2.3 Inteligencia Artificial en el derecho 21

2.3 Generalidades de procesamiento de lenguaje natural 22

2.3.1 Large Language Models (LLMs) 22

2.4 Inteligencia artificial en la gestión publica 25

2.4.1 Impacto de la Automatización en la Gestión Pública 25

2.4.2 Marco Legal 26

2.4.3 Marco Ético 27

2.5 Generalidades de procesos legales 27

2.5.1 Opinión Legal (Predictamen) 27

2.5.2 Normativas Legales 28

2.5.3 Automatización de Procesos Legales 29

2.6 Generalidades de procesos elaboración de opiniones legales 30

2.6.1 Ley N° 27444 - Ley del Procedimiento Administrativo General 30

2.7 Organización de la municipalidad provincial de Calca 30

2.7.1 Misión 31

2.7.2 Visión 31

2.7.3 Organigrama 32

2.7.4 Oficina de Asesoría Jurídica 33

2.8 Generalidades de estructura de modelos 34

2.8.1 Tokenización 34

2.8.2 Tokenización de Subpalabras 34

2.8.3 SFTTrainer 34

2.8.4 Input_ids 35

2.8.5 Attention_mask 35

2.8.6 Librerías – en general 35

2.9 Fine Tuning 40

2.9.1 Tipos de FineTuning 41

2.9.2 Overfitting 44

2.9.3 Training Loss 45

2.9.4 Validation Loss 45

2.10 Ingeniería de PROMPT 45

2.10.1 Clasificación de técnicas de prompting 47

Capítulo III: Marco metodológico 55

3.1 Tipo de la Investigación 55

3.2 Nivel de Investigación 55

3.3 Diseño metodológico de la investigación 56

3.3.1 Construcción de dataset jurídico 56

3.3.2 Métricas de evaluación 57

3.3.3 Comparación de modelos de lenguaje 58

3.3.4 Desarrollo del prototipo funcional 60

3.3.5 Evaluación de la calidad de las opiniones legales generadas 61

Capítulo IV: Desarrollo experimental del modelo IA 63

4.1 Selección de modelos LLM 63

4.2 Construcción de dataset jurídico 63

4.2.1 Dataset para Llama y Mistral 63

4.2.2 Dataset para Claude 64

4.3 Preprocesamiento de datos 65

4.3.1 Selección documental 65

4.3.2 Escaneo de opiniones legales 66

4.3.3 Digitalización y OCR 66

4.3.4 Conversión a formato JSON 66

4.3.5 Curación y limpieza de datos 67

4.3.6 Resultado 67

4.4 Entrenamiento de modelos Llama y Mistral 70
4.4.1 Tokenización 70
4.4.2 Adaptación del Modelo mediante LoRA 71
4.4.3 Fine-Tuning supervisado (SFT) con trl.SFTTrainer 72
4.4.4 Preparación del dataset para SFT 72
4.4.5 Configuración de SFTTrainer 73
4.4.6 Flujo de entrenamiento 75
4.4.7 Generación de Ejemplos 76
4.4.8 Guardado y Reutilización del Modelo 77
4.5 Técnicas de ingeniería de prompts aplicado a la investigación 77
4.5.1 Técnica Plan-and-Solve Prompting 78
4.5.2 Técnica Chain of Verification Prompting 83
4.6 Ejecución Experimental 86
4.6.1 Modelo Llama 3.1 88
4.6.2 Modelo Mistral 92
4.6.3 Modelo Claude 95
4.7 Evaluación de rendimiento 96
4.7.1 Criterio multicriterio para la selección del modelo 99
4.7.2 Normalización de los indicadores 100
4.7.3 Cálculo de las dimensiones 101

4.7.4 Cálculo del puntaje global 102

4.7.5 Resultados del análisis multicriterio 102

4.7.6 Definición de las dimensiones de evaluación 104

Capitulo V: Propuesta y diseño 106

5.1 Arquitectura del sistema para mostrar funcionalidad 106

5.1.1 Carga de documentos PDF 107

5.1.2 Procesamiento OCR 107

5.1.3 Extracción de antecedentes 108

5.1.4 Generación del informe legal 109

5.2 Componentes del sistema 110

5.2.1 Módulo de carga de archivos 110

5.2.2 Módulo OCR 111

5.2.3 Módulo de extracción de antecedentes 111

5.2.4 Módulo de generación del informe legal 111

5.2.5 Módulo de validación 112

5.3 Flujo funcional del sistema 112

5.4 Evaluación humana 113

5.4.1 Sustento de los criterios de evaluación 114

5.5 Tiempo total de atención de opiniones legales mediante el método tradicional 119

5.5.1 Determinación del tamaño de muestra de los expedientes escaneados 120

5.5.2 Resultados del tiempo de emisión de opiniones legales 121

5.6 Tiempo de elaboración de opiniones legales mediante el prototipo basado en inteligencia artificial 125

5.6.1 Etapas del proceso asistido mediante inteligencia artificial 126

5.7 Criterio para la comparación del tiempo de elaboración 126

5.7.1 Proceso tradicional 127

5.7.2 Prototipo basado en inteligencia artificial 127

5.7.3 Comparación de los indicadores temporales 128

5.7.4 Consideración metodológica 128

Capítulo VI: Resultados y Análisis 130

6.1 Resultados correspondientes al objetivo específico 1: Comparación de Modelos de lenguaje a gran escala 131

6.2 Resultados correspondientes al objetivo específico 2: Desarrollo del prototipo funcional 135

6.3 Resultados correspondientes al objetivo específico 3: Evaluación de la calidad de las opiniones legales generadas por el prototipo funcional 136

6.4 Discusión de resultados 141

Conclusiones 144

Recomendaciones 145

Trabajos a futuro 146

Bibliografía 147

Apéndice A 153

Índice de figuras

Figura 1. Organigrama de la municipalidad provincial de Calca 32

Figura 2. Diseño metodológico 56

Figura 3 Proceso de construcción dataset jurídico. Elaboración propia 65

Figura 4. Ejecución del proceso de entrenamiento del modelo Llama 3.1 75

Figura 5. Ejecución del proceso de entrenamiento del modelo 76

Figura 6. Métricas de calidad del modelo Llama 3.1 90

Figura 7. Métricas de calidad del modelo Mistral v0.3. 94

Figura 8. Tiempo promedio de inferencia del modelo Claude 96

Figura 9. Arquitectura del sistema para mostrar la funcionalidad 106

Figura 10. Interfaz del sistema que utiliza el abogado 106

Figura 11. Gráfico de barras por criterio del prototipo 139

Figura 12. Gráfico de radar del desempeño del prototipo 140

Figura 13. Gráfico de barras del promedio por especialista 140

Figura 14. Análisis comparativo de la evaluación entre el modelo Claude propuesto VS el modelo Claude de referencia 142

Índice de tablas

Tabla 1 Características más importantes de cada enfoque de Fine-Tuning..... 44

Tabla 2 Escala Likert de cinco niveles 61

Tabla 3 Comparación de modelos..... 98

Tabla 4 Asignación de pesos 99

Tabla 5 Resultados del análisis multicriterio..... 102

Tabla 6 Calidad y eficiencia demostrados según componentes 104

Tabla 7 Integración de criterios de evaluación..... 119

Tabla 8 Tiempo de emisión de opiniones legales mediante proceso manual..... 122

Tabla 9 Estadísticos descriptivos de tiempo de emisión de opiniones legales mediante procesos manual..... 124

Tabla 10 Etapas del proceso de elaboración de opiniones legales..... 126

Tabla 11. Comparación temporal del proceso tradicional y el prototipo..... 128

Tabla 12 Evaluación del prototipo por especialistas jurídicos 138

Capítulo I: Aspectos generales

1.1 Planteamiento de Problema

1.1.1 Descripción del problema

En los últimos años, la transformación digital y el avance de la inteligencia artificial han generado cambios significativos en la automatización de procesos documentales y administrativos en distintos sectores. Particularmente, los Modelos de lenguaje a gran escala (LLM) han demostrado capacidades relevantes para el procesamiento de información, generación automatizada de textos y asistencia en tareas especializadas que requieren análisis documental complejo.

En el ámbito jurídico y de la administración pública, el incremento constante de expedientes administrativos, documentos normativos y procedimientos legales ha generado mayores exigencias sobre las oficinas responsables de la emisión de opiniones legales. Esta situación demanda procesos de revisión exhaustivos, análisis normativo preciso y tiempos de respuesta oportunos, factores que frecuentemente se ven afectados por limitaciones operativas y alta carga laboral.

En el contexto peruano, diversas entidades públicas enfrentan restricciones presupuestales y limitaciones para la contratación de personal especializado. En particular, la Ley N.º 31298 establece restricciones relacionadas con modalidades de contratación temporal en el sector público. Asimismo, las políticas de austeridad implementadas en distintas instituciones del Estado han limitado la ampliación de recursos humanos, afectando la capacidad operativa de áreas estratégicas como las oficinas de asesoría jurídica.

En este contexto, la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca enfrenta importantes limitaciones operativas derivadas de la creciente carga laboral y la insuficiencia de personal especializado. El volumen de trabajo, conformado por la revisión de

leyes, decretos legislativos, directivas y demás documentos normativos, ha experimentado un incremento sostenido, lo que exige un análisis riguroso para garantizar la correcta emisión de opiniones legales. No obstante, la capacidad del personal disponible resulta insuficiente para atender dicha demanda, incrementando el riesgo de errores en la interpretación y aplicación de la normativa, con posibles repercusiones legales para la entidad.

Esta problemática se ve agravada por las restricciones normativas y presupuestales existentes. Pese a las reiteradas solicitudes de ampliación de personal formuladas mediante el INFORME N.º 029-2023-MPC-GM/OAJ y el INFORME N.º 135-2023-MPC-GM/OAJ, estas no han sido atendidas. A ello se suma el MEMORÁNDUM MÚLTIPLE N.º 0277-2023-GM-MPC/JLVG, el cual establece medidas de austeridad presupuestaria, restringiendo aún más la incorporación de nuevos profesionales.

Actualmente, la Oficina de Asesoría Jurídica atiende aproximadamente 100 expedientes administrativos mensuales, según registros internos, lo que evidencia una elevada carga operativa para el personal disponible. Asimismo, se ha identificado que el tiempo de emisión de las opiniones legales oscila entre 1 y 5 días por expediente, reflejando limitaciones en la capacidad de respuesta y variabilidad en los tiempos de atención.

Como consecuencia, se generan retrasos en la emisión de opiniones legales, acumulación de expedientes y dificultades en la revisión exhaustiva de antecedentes y normativa aplicable. Adicionalmente, la ausencia de herramientas tecnológicas orientadas al procesamiento automatizado de información jurídica limita la optimización de los procesos internos, afectando la eficiencia operativa de la Oficina de Asesoría Jurídica y el cumplimiento oportuno de sus funciones institucionales.

1.2 Formulación del problema

1.2.1 Problema general

¿En qué medida la implementación de una solución basada en inteligencia artificial generativa contribuirá a incrementar la eficiencia y la calidad del proceso de elaboración de opiniones legales en la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca?

1.2.2 Problemas específicos

- ¿De qué manera el uso de distintos Modelos de lenguaje a gran escala (LLM) influye en la calidad de la generación de opiniones legales en la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca?
- ¿En qué medida la automatización del procesamiento de expedientes administrativos mediante inteligencia artificial generativa contribuirá a reducir los tiempos de elaboración de opiniones legales?
- ¿De qué manera la implementación de un prototipo funcional basado en inteligencia artificial generativa influirá en la coherencia jurídica, precisión normativa y claridad de redacción de las opiniones legales generadas?

1.3 Objetivos

1.3.1 Objetivo general

Determinar en qué medida la implementación de una solución basada en inteligencia artificial generativa contribuye a incrementar la eficiencia y la calidad del proceso de elaboración de opiniones legales en la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca.

1.3.2 Objetivos específicos

- Analizar el desempeño de distintos Modelos de lenguaje a gran escala (LLM) en la generación de opiniones legales, a fin de identificar la alternativa más adecuada para el procesamiento de expedientes administrativos.
- Determinar la eficiencia del prototipo funcional basado en inteligencia artificial generativa en la automatización del procesamiento de expedientes y la generación de opiniones legales.
- Evaluar la calidad jurídica de las opiniones legales generadas por el sistema, considerando criterios de coherencia jurídica, precisión normativa, claridad de redacción y fundamentación argumentativa.

1.4 Justificación

La presente investigación se justifica por la necesidad de atender las limitaciones operativas que enfrenta la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca, derivadas del incremento de la carga documental, la insuficiencia de personal jurídico especializado y la complejidad de los procesos legales. En ese sentido, se exponen a continuación las principales dimensiones que sustentan la relevancia del trabajo de investigación.

1.4.1 Justificación tecnológica

Desde el punto de vista tecnológico, la investigación se justifica por la necesidad de incorporar herramientas innovadoras que permitan automatizar los procesos de revisión y redacción de documentos legales dentro de la Oficina de Asesoría Jurídica. En este contexto, se propone el uso de Inteligencia Artificial Generativa, específicamente mediante Modelos de lenguaje a gran escala (LLM), debido a su capacidad para procesar grandes volúmenes de

información textual, identificar antecedentes relevantes, analizar normativa aplicable y generar borradores de opiniones legales.

La elección de esta tecnología responde a su potencial para automatizar tareas repetitivas y de alto consumo de tiempo, tales como la revisión de leyes, decretos, informes previos y otros documentos administrativos. Asimismo, el uso de LLM podría contribuir a generar redacciones más precisas, coherentes y consistentes, favoreciendo la estandarización de los documentos emitidos por la oficina jurídica.

En consecuencia, esta investigación busca promover la incorporación de herramientas tecnológicas avanzadas en el ámbito jurídico municipal, contribuyendo al proceso de modernización administrativa y a la adopción de nuevas tecnologías en la gestión pública.

1.4.2 Justificación Teórica

Desde el punto de vista teórico, la investigación aporta al estudio de la aplicación de la Inteligencia Artificial en el ámbito del derecho administrativo y la gestión pública. Si bien existen diversos trabajos de investigación relacionados con el uso de IA en el sector privado, su aplicación en entidades públicas y, particularmente, en oficinas de asesoría jurídica municipal, continúa siendo limitada en el contexto peruano.

En ese sentido, la investigación contribuirá a ampliar el conocimiento sobre el uso de Modelos de lenguaje a gran escala en procesos de revisión de expedientes administrativos y redacción de opiniones legales, generando evidencia sobre sus alcances, limitaciones y posibilidades de implementación en el sector público. Asimismo, permitirá fortalecer la literatura vinculada al uso de tecnologías emergentes en la simplificación de procesos administrativos y jurídicos.

De igual manera, el presente trabajo de investigación servirá como antecedente para futuras investigaciones vinculadas a la automatización de funciones jurídicas, la modernización

de la gestión pública y la incorporación de herramientas de inteligencia artificial en las entidades del Estado.

1.4.3 Justificación Práctica

Desde una perspectiva práctica, la investigación se justifica porque busca proponer una solución viable frente a las limitaciones operativas de la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca. La implementación de una herramienta basada en Inteligencia Artificial contribuye a reducir los tiempos requeridos para la elaboración de opiniones legales, mejorar la organización de la carga de trabajo y optimizar el uso de los recursos humanos disponibles.

Asimismo, el sistema permitiría que los abogados concentren sus esfuerzos en actividades de mayor complejidad, tales como el análisis jurídico, la interpretación normativa y la toma de decisiones, dejando las tareas repetitivas y preliminares a cargo de la herramienta tecnológica.

En consecuencia, la investigación pretende demostrar que el uso de Inteligencia Artificial Generativa puede convertirse en una alternativa práctica para fortalecer la productividad institucional y mejorar la capacidad de respuesta de la oficina jurídica.

1.4.4 Justificación Metodológica

Desde el punto de vista metodológico, la investigación adquiere relevancia porque propone un esquema de evaluación cuantitativa del desempeño de sistemas basados en Inteligencia Artificial aplicados al ámbito jurídico. Para ello, se diseñará y aplicará una ficha de evaluación dirigida a abogados especialistas, la cual permitirá medir variables como coherencia jurídica, adecuación normativa, claridad de redacción, precisión en la interpretación de antecedentes, fundamentación argumentativa, utilidad práctica y nivel de autonomía del sistema.

Este enfoque permitirá obtener resultados medibles y comparables sobre el desempeño de los distintos modelos evaluados, generando evidencia objetiva respecto a su efectividad en la redacción de opiniones legales. Asimismo, el diseño metodológico podrá servir como referencia para futuras investigaciones relacionadas con la automatización de procesos jurídicos en entidades públicas.

1.4.5 Justificación Social

Desde el punto de vista social, la investigación es relevante porque la mejora de los procesos internos de la Oficina de Asesoría Jurídica repercute indirectamente en la ciudadanía. Una oficina jurídica con mayor capacidad de respuesta permite atender de manera más oportuna los procedimientos administrativos, reduciendo retrasos y fortaleciendo la eficiencia institucional.

Asimismo, la emisión de opiniones legales más precisas y oportunas contribuye a una mejor toma de decisiones dentro de la municipalidad, lo que beneficia a la población mediante una gestión pública más eficiente, transparente y orientada al cumplimiento de la normativa vigente.

En consecuencia, la investigación aporta al fortalecimiento de la confianza ciudadana en las instituciones públicas y promueve el uso responsable de tecnologías emergentes para mejorar los servicios que brinda la administración pública.

1.4.6 Justificación Legal

Desde una perspectiva legal, la investigación se justifica por la necesidad de optimizar los procesos de la Oficina de Asesoría Jurídica dentro del marco de las disposiciones normativas vigentes. Las restricciones presupuestales y administrativas existentes han limitado la posibilidad de ampliar el personal jurídico con experiencia, generando una mayor carga de trabajo y retrasos en la emisión de opiniones legales.

En este contexto, la incorporación de herramientas de Inteligencia Artificial se encuentra alineada con las disposiciones orientadas a la Ley N° 27658 - Ley marco de la modernización del estado, la Ley N° 25035 – Ley de simplificación administrativa y la Ley N.° 31814 - Ley que promueve el uso de la inteligencia artificial en favor del desarrollo económico y social del país. Asimismo, su implementación debe desarrollarse respetando los principios de legalidad, transparencia, confidencialidad y protección de datos.

La investigación no busca reemplazar la función del abogado, sino incorporar una herramienta de apoyo que opere bajo supervisión profesional y que contribuya a fortalecer la calidad de las opiniones legales emitidos por la institución.

1.4.7 Justificación Económica

Desde el punto de vista económico, la investigación se justifica porque la implementación de una herramienta basada en Inteligencia Artificial podría constituir una alternativa de optimización de recursos institucionales. Al automatizar determinadas tareas repetitivas, la oficina jurídica podría reducir la necesidad de destinar mayores recursos humanos y tiempo operativo a labores preliminares de revisión y redacción documental.

Asimismo, la reducción de errores, omisiones o retrasos en la emisión de opiniones legales podría contribuir a evitar costos derivados de observaciones, rectificaciones, controversias administrativas.

En consecuencia, la propuesta no solo busca mejorar el desempeño operativo de la oficina jurídica, sino también demostrar la viabilidad económica de incorporar tecnologías emergentes en contextos de restricción presupuestaria.

1.5 Alcances y limitaciones

1.5.1 Alcances

La presente investigación se desarrolla en la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca y se enfoca en la evaluación de un sistema basado en Inteligencia Artificial Generativa para la redacción automática de opiniones legales.

El trabajo de investigación comprende el análisis y comparación del desempeño de los modelos de lenguaje a gran escala como Claude, Llama y Mistral, aplicados a tareas de extracción de antecedentes, identificación de normativa y generación de predictamenes de opiniones legales. Asimismo, considera la aplicación de técnicas de ingeniería de prompts orientadas a mejorar la coherencia jurídica, la adecuación normativa y la estructura argumentativa de las opiniones legales generadas.

Finalmente, el presente trabajo de investigación busca determinar la viabilidad del uso de Inteligencia Artificial Generativa como herramienta de apoyo en la gestión jurídica municipal, sin sustituir el criterio profesional ni la responsabilidad funcional de los abogados.

1.5.2 Limitaciones

La investigación presenta algunas limitaciones que deben ser consideradas al momento de interpretar sus resultados.

En primer lugar, este trabajo de investigación se desarrolla únicamente en la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca, por lo que los resultados obtenidos corresponden a un contexto institucional específico.

Asimismo, la evaluación del sistema depende del criterio subjetivo de los abogados especialistas que participen en la valoración de los opiniones legales generados, lo cual podría generar diferencias en las calificaciones otorgadas.

Otra limitación importante radica en que los modelos de lenguaje a gran escala utilizados pueden presentar errores en la interpretación de antecedentes, omisiones normativas o generación de información inexacta, por lo que los resultados producidos por el sistema siempre requieren supervisión y validación humana.

Finalmente, la investigación se centra únicamente en opiniones legales de naturaleza administrativa, sin abarcar otras ramas del derecho como materia penal, civil o procesal.

Capítulo II: Marco teórico

2.1 Antecedentes de la Investigación

2.1.1 Antecedentes Internacionales

(Fuentes de Abreu, 2025) Generación de Resúmenes Judiciales utilizando Modelos de Lenguaje (LLMs)

La presente investigación tuvo como objetivo desarrollar y evaluar un sistema basado en modelos de lenguaje de gran tamaño (LLMs) para la generación automática de resúmenes de documentos judiciales, analizando la calidad de los resúmenes producidos respecto a los elaborados manualmente.

En la investigación se empleó un enfoque experimental basado en documentos judiciales proporcionados por el Ministerio de Justicia de la Ciudad de Córdoba. El conjunto de datos estuvo conformado por 470 documentos, organizados en categorías civiles y penales. Los documentos fueron preprocesados mediante PyMuPDF, NLTK y spaCy, aplicando técnicas de limpieza, tokenización y clustering. Posteriormente, se implementó una arquitectura basada en prompt engineering para generar resúmenes utilizando diferentes modelos de lenguaje, sin realizar procesos de fine-tuning.

Para evaluar el desempeño del sistema, los autores compararan los resúmenes generados por los modelos de lenguaje con los resúmenes elaborados manualmente, analizando criterios como coherencia, precisión y fidelidad del contenido, además de comparar el rendimiento entre distintos modelos de lenguaje utilizados durante los experimentos.

En conclusión, los resultados evidenciaron que los modelos de lenguaje fueron capaces de generar resúmenes jurídicos coherentes y útiles como borradores iniciales, aunque presentaron limitaciones relacionadas con la precisión factual y el manejo de documentos de

gran longitud. Asimismo, se observaron diferencias de desempeño entre los modelos evaluados, dependiendo de la estrategia de prompting empleada.

(Ashley & Phelps, 2022) "ALEXA, WRITE A MEMO: THE PROMISE AND CHALLENGES OF AIAND LEGAL WRITING" Washington DC, EEUU, 2022.

La presente investigación analiza el potencial y las limitaciones de la inteligencia artificial (IA) en la redacción legal, en particular en la creación de memorandos de oficina. Aunque la IA ya ayuda en tareas como la revisión de contratos, la investigación jurídica y la predicción de resultados mediante aprendizaje automático y análisis de texto, aún enfrenta grandes desafíos en áreas como la interpretación de leyes y la redacción de argumentos jurídicos complejos. Los autores destacan que, aunque la IA puede automatizar algunos aspectos de la escritura legal, los abogados siguen siendo esenciales para garantizar la precisión y el razonamiento jurídico. Además, sugieren que las facultades de derecho deben adaptarse para enseñar a los estudiantes a trabajar en colaboración con herramientas de IA, reconociendo que el pensamiento crítico y las decisiones humanas siguen siendo indispensables en la práctica legal.

Dentro de la Metodología se utilizó un enfoque teórico y comparativo para analizar el estado actual de la inteligencia artificial en la redacción legal. A través de la revisión de herramientas como Westlaw Edge y Casetext, los autores evalúan qué tareas legales, como la investigación y predicción de casos, ya pueden ser automatizadas por IA, y cuáles, como la redacción analítica y la interpretación legal, aún dependen del razonamiento humano. Además, el artículo sugiere cómo las facultades de derecho deben preparar a los estudiantes para colaborar con estas tecnologías en su futura práctica profesional.

Las métricas utilizadas no tienen una especifica, pero se tiene la Precisión, velocidad y capacidad de automatización de la IA en la elaboración de Memorándums considerando la búsqueda de documentos legales

La conclusión en la implementación de la IA en la redacción de Memorándums es que, aunque la inteligencia artificial tiene el potencial de automatizar ciertas tareas legales como la investigación y el análisis de textos, todavía enfrenta limitaciones significativas en la redacción de documentos analíticos complejos y la interpretación de leyes. Por lo tanto, los abogados siguen siendo esenciales para garantizar precisión y razonamiento jurídico. El artículo también resalta la necesidad de que las facultades de derecho adapten su enseñanza para preparar a los estudiantes a colaborar con herramientas de IA en la práctica profesional.

(Mongoli, 2024)"THE USE OF LLMS IN THE LEGAL FIELD: OPTIMIZING CONTRACT MANAGEMENT WITH GENERATIVE ARTIFICIAL INTELLIGENCE" Universidad Politecnico di Torino, Italia.

Esta investigación explora la aplicación de Modelos de Lenguaje de Gran Escala (LLMs) en el ámbito legal, específicamente para optimizar la gestión de contratos mediante la inteligencia artificial generativa. Se desarrolla un prototipo de concepto (POC) en forma de aplicación web, que permite a los abogados analizar y generar contratos de manera más eficiente. El trabajo de investigación combina técnicas de Recuperación Aumentada por Generación (RAG) y ajustes de prompts para mejorar la precisión y rapidez en la generación de cláusulas contractuales.

Se empleó un enfoque basado en LLMs como GPT-3 para la generación de contratos y el análisis de cláusulas. La metodología incluyó la implementación de un sistema RAG para mejorar la generación de textos mediante la recuperación de información relevante de una base de datos. La aplicación se desarrolló con Streamlit para el front-end y Redis para almacenar y

gestionar datos vectorizados, lo que permitió una integración eficiente entre los modelos de IA y el sistema de recuperación de información.

Las métricas utilizadas incluyen la eficiencia en el tiempo de respuesta y la precisión en la generación de cláusulas contractuales. Se comparó la precisión de las cláusulas generadas automáticamente con las creadas manualmente por abogados, midiendo la reducción de errores humanos y el tiempo total de procesamiento. También se evaluó la satisfacción de los usuarios con el sistema, midiendo la facilidad de uso y la adopción de la herramienta.

La implementación de LLMs en el ámbito legal mostró resultados prometedores al mejorar significativamente la eficiencia en la gestión de contratos. La combinación de RAG y ajuste de prompts permitió la generación de cláusulas con mayor precisión y rapidez. Aunque el sistema necesita supervisión humana para asegurar la coherencia legal, se demostró que la IA puede reducir la carga de trabajo repetitiva y mejorar la productividad general en el sector legal.

(Hadi, y otros, 2023), “A COMPREHENSIVE SURVEY OF APPLICATIONS, CHALLENGES, DATASETS, LIMITATIONS, AND FUTURE PROSPECTS” arXiv.

El presente articulo ofrece un análisis exhaustivo sobre los Modelos de lenguaje a gran escala (LLM) desde 2015 en adelante, destacando su impacto en el procesamiento del lenguaje natural (NLP). Los LLM, especialmente los basados en arquitecturas como transformadores y GPT, han mejorado significativamente la capacidad de generar y comprender texto a nivel casi humano, con aplicaciones en áreas como medicina, educación, derecho e ingeniería. Sin embargo, también enfrentan desafíos importantes, como cuestiones éticas, precisión, y preocupaciones relacionadas con la seguridad y privacidad. El artículo subraya tanto los beneficios como las limitaciones de los LLM y plantea la necesidad de investigaciones futuras para abordar estos retos y mejorar su rendimiento.

La metodología del documento se centra en una revisión exhaustiva de los Modelos de lenguaje a gran escala (LLMs), recopilando y analizando investigaciones previas, arquitecturas clave como Transformers, y sus aplicaciones en diferentes campos. Se destacan las capacidades, limitaciones y desafíos éticos de los LLMs, proponiendo pautas para su uso futuro y mejorado en la sociedad.

La investigación utiliza varias métricas para evaluar el rendimiento de los Modelos de lenguaje a gran escala (LLM), como la inspección visual, que permite una evaluación subjetiva de la calidad por expertos, y el Inception Score (IS), que mide tanto la calidad como la diversidad de las muestras generadas. También se utiliza la métrica Frechet Inception Distance (FID), que compara la similitud entre las distribuciones de las muestras reales y generadas. Otras métricas mencionadas incluyen curvas de precisión y recuperación, estudios de usuarios, y métricas de cobertura, con el fin de obtener una evaluación integral del desempeño de los LLM.

La Investigación proporciona una visión integral sobre los Modelos de lenguaje a gran escala (LLM), explorando su evolución, implicaciones técnicas, y usos prácticos. Se destacan tanto los beneficios como los riesgos de los LLM, con ejemplos de su aplicación en áreas como la medicina, la ingeniería, la educación y el derecho. Además, se presentan directrices para su uso óptimo y responsable, subrayando la importancia de abordar las preocupaciones éticas y de seguridad. El documento sugiere que, aunque los LLM han avanzado significativamente, aún hay desafíos que deben ser resueltos para maximizar su potencial en beneficio de la sociedad.

Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, *36*, 46595-46623.

Este trabajo de investigación se centra en evaluar el uso de modelos de lenguaje grande (LLM) como jueces para la evaluación de asistentes de chat, particularmente en escenarios de preguntas abiertas y múltiples turnos. Utilizando dos enfoques, MT-bench y Chatbot Arena, se analizan las preferencias humanas en interacciones con chatbots, identificando cómo los LLM alineados con el juicio humano pueden servir como herramientas efectivas para la evaluación automática.

La investigación emplea MT-bench, un conjunto de preguntas que evalúa las capacidades de los chatbots en conversaciones de múltiples turnos. También se utiliza Chatbot Arena, una plataforma de evaluación basada en crowdsourcing donde los usuarios comparan dos chatbots simultáneamente. Los datos obtenidos son comparados con evaluaciones de jueces humanos para validar la eficacia de los LLM como evaluadores.

Se mide el grado de acuerdo entre los jueces LLM y humanos, con un énfasis en superar el 80% de acuerdo en los juicios. Además, se analizan métricas como la tasa de victoria de los modelos y la consistencia del juicio en diferentes categorías de preguntas, desde matemáticas hasta habilidades conversacionales.

Los resultados indican que los LLM como GPT-4 pueden igualar el nivel de acuerdo entre jueces humanos, haciendo de los LLM una herramienta escalable y eficiente para la evaluación de chatbots. Sin embargo, se reconocen limitaciones como los sesgos de posición y la capacidad limitada en problemas matemáticos complejos, lo que sugiere áreas para futuras mejoras.

2.1.2 Antecedentes Nacionales

(Doroteo Guerrero & Camacho Ortiz, 2021), "INTELIGENCIA ARTIFICIAL Y LA AUTOMATIZACIÓN DE PROCESOS JUDICIALES EN LA ADMINISTRACIÓN DE JUSTICIA DEL PERÚ, 2021" Universidad Privada del Norte, Lima.

Esta investigación tiene como objetivo explorar cómo la inteligencia artificial (IA) puede integrarse en los procesos judiciales para mejorar la eficiencia y la rapidez en la administración de justicia en el Perú. El trabajo de investigación emplea un método sociológico con un enfoque cualitativo y un diseño de estudio de casos. Investiga cómo se podría utilizar la IA, particularmente en casos de derecho de familia como los de pensión alimenticia, para automatizar las resoluciones judiciales. Al implementar la IA, el trabajo de investigación concluye que es posible reducir el tiempo y el costo asociados con los procedimientos legales, manteniendo la equidad y precisión. Los resultados sugieren que la IA tiene un impacto favorable en la automatización de los procesos judiciales, siendo una herramienta necesaria para mejorar la administración de justicia en el Perú.

La metodología utilizada en la investigación tiene un enfoque cualitativo y se clasifica como un estudio de caso. Las investigadoras emplearon un análisis documental y entrevistas a abogados especialistas en Derecho de Familia y nuevas tecnologías de IA. El diseño del trabajo de investigación permite analizar cómo la IA puede mejorar la eficiencia en los procesos judiciales. Se realizaron entrevistas no estructuradas a 8 abogados (6 nacionales y 2 internacionales) a través de la plataforma Zoom. El procedimiento incluyó la recolección de datos de fuentes documentales relevantes y la posterior transcripción y análisis de las entrevistas. Las investigadoras validaron las respuestas y el material recopilado siguiendo parámetros académicos establecidos.

Las métricas usadas dentro de la investigación se centran en las mejoras de efectividad y eficiencia observadas al usar IA, específicamente el sistema Prometea, en los procesos judiciales. Las métricas principales incluyen:

Reducción de Tiempo: Prometea demostró una reducción significativa del tiempo en varios procesos judiciales. Por ejemplo, el tiempo para crear un decreto de hechos se redujo de 25 minutos a 2 minutos, y la revisión de contratos pasó de 2 horas a solo 1 minuto.

Aumento en la Precisión**:** Prometea redujo los errores tipográficos en un 99%, mejorando la precisión de los documentos.

Mejoras en la Eficiencia**:** En algunos procesos, el tiempo requerido para tareas legales específicas se redujo hasta en un 80%, mostrando que la IA acelera significativamente las tareas que tradicionalmente consumen mucho tiempo

Las conclusiones de la tesis destacan que la implementación de la inteligencia artificial (IA) en el sistema judicial peruano tiene un impacto positivo al automatizar los procesos judiciales, especialmente en el Derecho de Familia. La IA mejora la eficiencia, reduce significativamente los tiempos procesales y aumenta la predictibilidad de los resultados, proporcionando respuestas precisas y personalizadas basadas en algoritmos y datos previos. Además, se plantea como una solución alternativa viable para superar las limitaciones actuales del sistema judicial tradicional, optimizando el cumplimiento de los plazos procesales y mejorando la administración de justicia en el Perú.

(Ramirez Martinez, 2021). Asistente virtual para apoyar la atención al cliente en consultas legales de ámbito laboral en la empresa abogados Romero.

La presente investigación tiene como objetivo mejorar la atención al cliente en la empresa Abogados Romero mediante la implementación de un asistente virtual basado en inteligencia artificial. Este asistente virtual busca resolver consultas frecuentes de ámbito legal,

optimizando el tiempo de respuesta y mejorando la experiencia del usuario. A nivel internacional, los chatbots se han adoptado con éxito para atender diversas interacciones con los clientes, y esta investigación busca explorar su aplicación en el campo jurídico, particularmente en el derecho laboral.

Se utiliza una metodología CommonKADS para el desarrollo del sistema basado en conocimientos, que incluye varios modelos como el de organización, tareas, agentes, comunicaciones, conocimiento y diseño. Se optó por esta metodología debido a su capacidad para integrar sistemas inteligentes, lo que permitió estructurar el asistente virtual de manera eficiente para cubrir las necesidades de los clientes de la empresa.

Las métricas utilizadas incluyen el número de consultas resueltas por el asistente virtual, la reducción en el tiempo de respuesta a las consultas, la tasa de satisfacción de los usuarios con las respuestas proporcionadas y la efectividad en la automatización de tareas repetitivas. Se aplicarán pre-tests y post-tests para medir las mejoras en la atención al cliente antes y después de implementar el asistente virtual.

La implementación de un asistente virtual ha demostrado ser eficaz en la mejora de la atención al cliente en la empresa Abogados Romero. Se logró una reducción significativa en los tiempos de respuesta y un aumento en la satisfacción de los clientes. Además, el sistema ha permitido optimizar las consultas repetitivas, liberando tiempo para que los abogados se concentren en asuntos más complejos.

2.2 Bases Teóricas

2.2.1 Inteligencia artificial (IA)

La Inteligencia Artificial (IA) se refiere a la capacidad de un sistema computacional para realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, el razonamiento, la toma de decisiones, y la comprensión del lenguaje natural. La IA emplea

algoritmos avanzados y modelos de aprendizaje automático para procesar grandes cantidades de datos, identificar patrones, y realizar predicciones o decisiones de manera autónoma.

2.2.2 Inteligencia Artificial en la Administración Pública

La Inteligencia Artificial (IA) en la Administración Pública ha emergido como una herramienta fundamental para la modernización y mejora de los procesos gubernamentales. La implementación de IA en este ámbito no solo busca optimizar la eficiencia operativa y la toma de decisiones, sino también aumentar la transparencia y la calidad de los servicios públicos. En particular, la automatización de procesos legales mediante IA se ha destacado como un área de creciente importancia.

(López & Silva, Transformación Digital en la Administración Pública: El Impacto de la IA, 2021) señalan que la IA tiene el potencial de transformar radicalmente la administración pública, permitiendo que las tareas rutinarias y repetitivas sean automatizadas, lo que libera a los funcionarios para que se concentren en actividades de mayor valor añadido. Esta automatización incluye desde la gestión de trámites administrativos hasta la revisión y redacción de documentos legales, donde la precisión y la eficiencia son cruciales.

En el contexto legal, (García Marín, 2020) destaca que la automatización mediante IA puede mejorar significativamente la revisión de contratos y otros documentos legales, reduciendo el tiempo necesario para estas tareas y minimizando los errores humanos. La IA permite la identificación automática de cláusulas problemáticas, la estandarización de términos legales, y la redacción coherente y conforme a las normativas vigentes.

Por otro lado, (Robles & Martínez, 2022) argumentan que la adopción de IA en la administración pública también presenta desafíos importantes, como la necesidad de asegurar la equidad en la toma de decisiones automatizadas y la protección de datos personales. Estos

desafíos deben ser abordados mediante marcos normativos adecuados que garanticen un uso ético y responsable de la tecnología.

Finalmente, (Martín & Hernández, Capacitación en Nuevas Tecnologías para la Administración Pública, 2020) subrayan la importancia de la capacitación continua del personal para asegurar una implementación exitosa de la IA en la administración pública. Sin una formación adecuada, existe el riesgo de que las tecnologías de IA no se utilicen de manera óptima, lo que podría limitar sus beneficios potenciales.

2.2.3 Inteligencia Artificial en el derecho

Inteligencia Artificial (IA) en el derecho se refiere a la aplicación de tecnologías avanzadas de IA en el ámbito jurídico para mejorar la eficiencia, precisión, y accesibilidad de los servicios legales. La IA en el derecho abarca un amplio rango de tecnologías y aplicaciones, incluyendo el uso de algoritmos de aprendizaje automático (machine learning), procesamiento de lenguaje natural (NLP), análisis predictivo, y sistemas expertos, entre otros, para asistir a abogados, jueces y otros profesionales legales en la toma de decisiones, la gestión de casos, la redacción de documentos, y la investigación jurídica.

La IA permite la automatización de tareas repetitivas y de alto volumen en la práctica legal, como la revisión de documentos, la clasificación de casos, y la gestión de calendarios judiciales. Además, con la ayuda del procesamiento de lenguaje natural (NLP), los sistemas de IA pueden analizar grandes volúmenes de datos legales, incluidos casos judiciales, leyes y regulaciones, para identificar patrones, precedentes y referencias relevantes. También, los algoritmos de aprendizaje automático pueden ser utilizados para predecir el resultado probable de casos judiciales basándose en datos históricos, lo que ayuda a los abogados a evaluar la viabilidad de un caso y determinar estrategias legales más efectivas.

2.3 Generalidades de procesamiento de lenguaje natural

2.3.1 Large Language Models (LLMs)

Los Large Language Models (LLMs) son modelos de inteligencia artificial (IA) de gran escala diseñados para comprender, generar, y manipular texto en lenguaje natural. Estos modelos se basan en la arquitectura de transformadores, introducida por (Vaswani, y otros, 2017), y están preentrenados con grandes cantidades de datos textuales, lo que les permite generar respuestas coherentes y relevantes a partir de un contexto dado. Los LLMs, como GPT-3 de OpenAI y BERT de Google, son capaces de realizar una amplia gama de tareas de procesamiento de lenguaje natural (NLP), incluyendo la traducción de idiomas, la redacción de ensayos, la respuesta a preguntas, y el análisis de sentimientos.

2.3.1.1 Definición y funcionamiento

Según (Brown & Amodei, 2020), los LLMs "se entrenan utilizando billones de parámetros y grandes cantidades de texto no supervisado, lo que les permite desarrollar una representación profunda del lenguaje humano". Los LLMs están diseñados para predecir la próxima palabra en una secuencia de texto y, a través de este mecanismo, son capaces de generar oraciones, párrafos, o documentos completos que son coherentes con el contexto anterior.

Los LLMs funcionan de manera similar a los modelos tradicionales de NLP, pero con la diferencia de que tienen una mayor capacidad de generalización debido a su tamaño y complejidad. Los LLMs están preentrenados en vastos conjuntos de datos y luego pueden ser ajustados (fine-tuned) para tareas específicas utilizando datos adicionales.

2.3.1.2 Aplicaciones

Redacción automática de textos: Generación de artículos, ensayos, y descripciones de productos de manera automática.

Traducción de idiomas: Los LLMs, como BERT y GPT, se utilizan en sistemas de traducción automática de idiomas, mejorando la precisión y naturalidad de las traducciones.

Análisis de sentimientos y clasificación de texto: Se utilizan en aplicaciones que analizan el tono emocional de textos, como comentarios de redes sociales o reseñas de productos.

Asistentes virtuales: Los LLMs son la base de asistentes como Siri y Alexa, que generan respuestas naturales a preguntas complejas.

2.3.1.3 Funcionamiento del LLMs

El funcionamiento de un Large Language Model (LLM) se basa en la arquitectura de transformadores, diseñada para procesar grandes volúmenes de texto y generar respuestas coherentes basadas en patrones aprendidos. Aquí se explica cómo funciona:

El modelo LLM se entrena en grandes cantidades de texto, como libros, artículos, y otros datos no estructurados. Durante el preentrenamiento, el objetivo del modelo es predecir la siguiente palabra en una secuencia dada, utilizando técnicas como el *Masked Language Modeling (MLM)* (como en BERT) o el modelado causal (como en GPT).

(Vaswani, y otros, 2017)**,** en su trabajo sobre transformadores, explican que esta arquitectura permite que el modelo analice todas las palabras de una secuencia de manera simultánea y determine cuáles son las más relevantes mediante un mecanismo de atención. La atención es clave para que el modelo se enfoque en las partes más significativas del texto en un contexto dado.

Después del preentrenamiento, los LLMs pueden ser ajustados finamente para tareas específicas utilizando un conjunto de datos etiquetados. Esta fase ayuda a especializar al modelo en tareas como clasificación de texto, generación de lenguaje, traducción automática, etc.

El ajuste fino permite que el modelo se adapte a diferentes dominios, como el análisis de sentimientos o la revisión de contratos legales.

Durante la fase de generación, el modelo toma una entrada de texto (puede ser una pregunta o un contexto) y utiliza la representación interna que ha aprendido para predecir la palabra siguiente, generando así oraciones y párrafos coherentes.

Este proceso implica una predicción secuencial, donde el modelo genera palabra por palabra basándose en el contexto proporcionado y el conocimiento previo.

Los LLMs utilizan el mecanismo de atención para asignar diferentes pesos a cada palabra del contexto de entrada, permitiendo que el modelo se "centre" en las palabras más relevantes al generar una respuesta.

(Vaswani, y otros, 2017) describen el mecanismo de atención como la clave para permitir que los LLMs comprendan dependencias a largo plazo entre palabras que no están necesariamente cerca una de la otra en una oración.

Uno de los aspectos importantes de los LLMs es su escalabilidad. Al aumentar la cantidad de parámetros (hasta miles de millones en modelos como GPT-3), el modelo puede mejorar su capacidad para aprender patrones complejos del lenguaje. Sin embargo, esto también implica una mayor demanda de recursos computacionales para entrenar y desplegar estos modelos.

Una vez que el modelo está preentrenado y ajustado, puede usarse en diversas aplicaciones, como asistentes virtuales, chatbots, traducción automática y generación de contenido. Los LLMs generan texto en tiempo real, proporcionando respuestas inmediatas basadas en el input del usuario.

2.4 Inteligencia artificial en la gestión publica

2.4.1 Impacto de la Automatización en la Gestión Pública

E La Automatización en la Gestión Pública se refiere a la implementación de tecnologías avanzadas que permiten a las instituciones gubernamentales mejorar sus procesos, aumentar la eficiencia operativa y automatizar la toma de decisiones. La automatización incluye desde la digitalización de trámites y servicios hasta el uso de sistemas de inteligencia artificial para análisis de datos y apoyo en la toma de decisiones estratégicas. Esta tendencia ha demostrado ser un motor clave para la modernización del sector público.

2.4.1.1 Impacto en la eficiencia operativa y la toma de decisiones

2.4.1.1.1 Mejora de la Eficiencia Operativa:

Según (López & Fernández, 2021) , la automatización en la gestión pública ha permitido reducir significativamente los tiempos de respuesta en la administración de servicios públicos. Un ejemplo destacado es el uso de sistemas automatizados en las oficinas tributarias para el procesamiento de declaraciones fiscales, lo que ha resultado en una mayor eficiencia y una reducción de los errores humanos.

2.4.1.1.2 Automatización de la Toma de Decisiones:

La implementación de tecnologías automatizadas también ha mejorado la capacidad de las instituciones para tomar decisiones informadas y basadas en datos. (Robles & Martínez, 2022) mencionan el caso de la ciudad de Barcelona, donde el uso de inteligencia artificial para analizar datos de movilidad ha permitido una gestión más eficaz del tráfico y del transporte público, mejorando tanto la eficiencia como la sostenibilidad de las decisiones tomadas.Marco Legal y Ético del Uso de IA

El uso de la Inteligencia Artificial (IA) en los procesos legales plantea una serie de retos legales y éticos que deben ser considerados para asegurar su implementación adecuada en la

Oficina de Asesoría Jurídica del Municipio de Calca. Este marco teórico abarca tanto las normativas que regulan el uso de IA en el contexto legal, como las consideraciones éticas vinculadas al manejo de datos sensibles y la toma de decisiones automatizadas.

2.4.2 Marco Legal

El marco legal relacionado con el uso de la IA en procesos jurídicos abarca normativas nacionales e internacionales que establecen las directrices para su implementación y operación. En Perú, el uso de IA en el ámbito legal debe alinearse con las leyes vigentes sobre protección de datos, transparencia y responsabilidad.

2.4.2.1 Ley N.º 29733, Ley de protección de datos personales

Esta ley peruana regula el tratamiento de los datos personales y establece que cualquier sistema automatizado que procese datos de personas, como el sistema basado en la API de ChatGPT propuesto en este proyecto, debe garantizar la protección y privacidad de esos datos. En el contexto legal, esto es particularmente importante, dado que los documentos revisados y generados por la IA pueden contener información sensible.

El sistema propuesto debe cumplir con los principios de seguridad de la información, confidencialidad, integridad y disponibilidad de los datos, evitando accesos no autorizados o la manipulación de información jurídica.

2.4.2.2 Decreto Legislativo N.º 1246, Simplificación Administrativa

Este decreto promueve la simplificación de trámites en las entidades públicas y el uso de tecnologías para mejorar la eficiencia. La implementación de IA en la Oficina de Asesoría Jurídica se alinea con este marco normativo, ya que busca automatizar y optimizar los procesos de revisión y redacción de opiniones legales.

Este marco establece un contexto propicio para el uso de herramientas automatizadas, siempre que éstas respeten los principios de legalidad, transparencia y cumplimiento normativo.

2.4.2.3 Normativa Internacional sobre el Uso de IA

A nivel internacional, organismos como la Unión Europea han establecido marcos normativos para regular el uso de la IA, destacando principios como la responsabilidad, explicabilidad y la ética en la toma de decisiones automatizadas. La Carta de Derechos Fundamentales de la Unión Europea establece que las personas tienen derecho a una protección adecuada en caso de que sus datos sean procesados por sistemas automatizados.

Además, los Principios Éticos de la IA de la UNESCO destacan la importancia de la justicia, equidad y no discriminación en el uso de tecnologías de IA, asegurando que no se utilicen para tomar decisiones sesgadas o desiguales.

2.4.3 Marco Ético

El uso de IA en los procesos legales plantea dilemas éticos que deben ser abordados para asegurar una implementación responsable y justa. A continuación, se abordan las principales consideraciones éticas relacionadas con el uso de IA en la Oficina de Asesoría Jurídica.

2.5 Generalidades de procesos legales

2.5.1 Opinión Legal (Predictamen)

Una Opinión Legal, también conocida como predictamen, es un documento formal que expresa la interpretación legal de un abogado o asesor jurídico sobre un tema o situación específica. Este documento tiene como objetivo proporcionar una guía clara y fundamentada sobre cómo se deben interpretar y aplicar las leyes en un caso particular, ayudando a las partes interesadas a tomar decisiones informadas y a garantizar que sus acciones cumplan con las normativas vigentes.

Según (García & Martínez, 2021), la precisión en la redacción de una opinión legal es crucial porque cualquier error o ambigüedad puede dar lugar a malentendidos, disputas legales,

o incluso a la nulidad de un acto administrativo. La claridad es igualmente importante, ya que permite que todas las partes comprendan plenamente el contenido y las implicaciones de la opinión legal emitida.

(López & Hernández, 2022) subrayan que la elaboración de un predictamen requiere un análisis exhaustivo de la normativa aplicable, precedentes judiciales relevantes y cualquier otro factor legal que pueda influir en la interpretación del tema en cuestión. Este análisis debe estar cuidadosamente documentado y expresado en un lenguaje claro y preciso para evitar posibles interpretaciones erróneas.

(Martínez & Pérez, 2023) añaden que, en el contexto de la automatización de procesos legales, la precisión y claridad en las opiniones legales son aún más críticas. La automatización de la redacción de predictámenes mediante inteligencia artificial requiere que los sistemas estén bien entrenados y configurados para garantizar que los documentos generados cumplan con los altos estándares de precisión y claridad necesarios en el ámbito legal.

2.5.2 Normativas Legales

Las Normativas Legales son el conjunto de leyes, decretos, reglamentos y otras disposiciones legales que establecen el marco jurídico dentro del cual deben operar las instituciones públicas. Estas normativas tienen como objetivo garantizar que las acciones y decisiones de las administraciones públicas se realicen conforme a los principios de legalidad, transparencia, y eficiencia, protegiendo los derechos de los ciudadanos y asegurando el cumplimiento de los deberes del Estado.

Según (García & López, 2021), las normativas legales incluyen desde la Constitución, que establece los principios fundamentales del ordenamiento jurídico, hasta los decretos legislativos y reglamentos específicos que regulan áreas concretas de la administración pública. Estas normativas son esenciales para guiar la automatización de procesos legales, ya que

aseguran que los sistemas automatizados operen dentro de los límites legales y produzcan documentos que sean válidos y conformes a la ley.

(Martínez & Pérez, 2022) enfatizan que, en el contexto de la automatización de la revisión y redacción de documentos legales, es crucial que los sistemas de inteligencia artificial utilizados estén configurados para cumplir con todas las normativas aplicables. Esto incluye la correcta interpretación y aplicación de leyes nacionales, reglamentos específicos y directivas internas de las instituciones. Sin este cumplimiento, los documentos generados podrían ser impugnados o considerados nulos, lo que podría tener graves consecuencias legales.

(López & García, 2023) también destacan la importancia de los marcos normativos específicos, como los decretos legislativos que regulan la simplificación de trámites administrativos y la modernización del sector público. Estos marcos legales no solo establecen los requisitos que deben cumplir los documentos legales, sino que también fomentan la adopción de tecnologías avanzadas, como la inteligencia artificial, para mejorar la eficiencia y la calidad de los servicios públicos.

2.5.3 Automatización de Procesos Legales

La Automatización de Procesos Legales implica el uso de tecnologías avanzadas, como la inteligencia artificial (IA) y el aprendizaje automático, para realizar tareas legales rutinarias de manera más eficiente y precisa. Estas tareas incluyen la revisión de contratos, la redacción de documentos legales, y la gestión de casos, que tradicionalmente requieren mucho tiempo y esfuerzo manual. La automatización permite que estas tareas se realicen de manera más rápida, reduciendo los errores humanos y liberando a los profesionales del derecho para que se concentren en actividades más estratégicas y complejas.

Además, (Pérez & Hernández, 2023) destacan que la automatización de procesos legales también permite una mayor consistencia en los documentos generados, ya que los sistemas

automatizados aplican las mismas reglas y criterios a cada documento, eliminando las variaciones que podrían surgir en la redacción manual. Esto es particularmente útil en grandes organizaciones o instituciones públicas donde es crucial mantener un alto nivel de uniformidad en los documentos legales.

2.6 Generalidades de procesos elaboración de opiniones legales

El trabajo de un asesor legal en Perú está regulado por un conjunto de normas y principios que aseguran que su desempeño sea conforme a la ley, ético, y profesional. A continuación, se describen las principales normas y regulaciones que rigen el trabajo de un asesor legal:

2.6.1 Ley N° 27444 - Ley del Procedimiento Administrativo General

Ámbito: Aplica cuando el asesor legal trabaja dentro de una entidad pública, regulando los procedimientos administrativos en los que puede estar involucrado.

Implicaciones: Los asesores legales deben asegurarse de que los procedimientos administrativos sean transparentes, respetuosos de los derechos de los administrados, y conformes a los principios de legalidad y debido proceso.

2.7 Organización de la municipalidad provincial de Calca

La Municipalidad Provincial de Calca es el órgano de gobierno local, emanado por la voluntad popular, representa al vecindario, promueve la adecuada prestación de servicios públicos locales y el desarrollo económico social local, con personería jurídica de derecho púbico y plena capacidad para el cumplimiento de sus fines. Goza de autonomía, política económica y administrativa en los asuntos de su competencia con sujeción al ordenamiento jurídico conforme lo establece la Constitución Política del Estado, la Ley Orgánica de Municipalidades y demás normas legales vigentes a la fecha. Ejerce la administración municipal del distrito Capital, extendiendo su jurisdicción al ámbito de la Provincia de Calca.

2.7.1 Misión

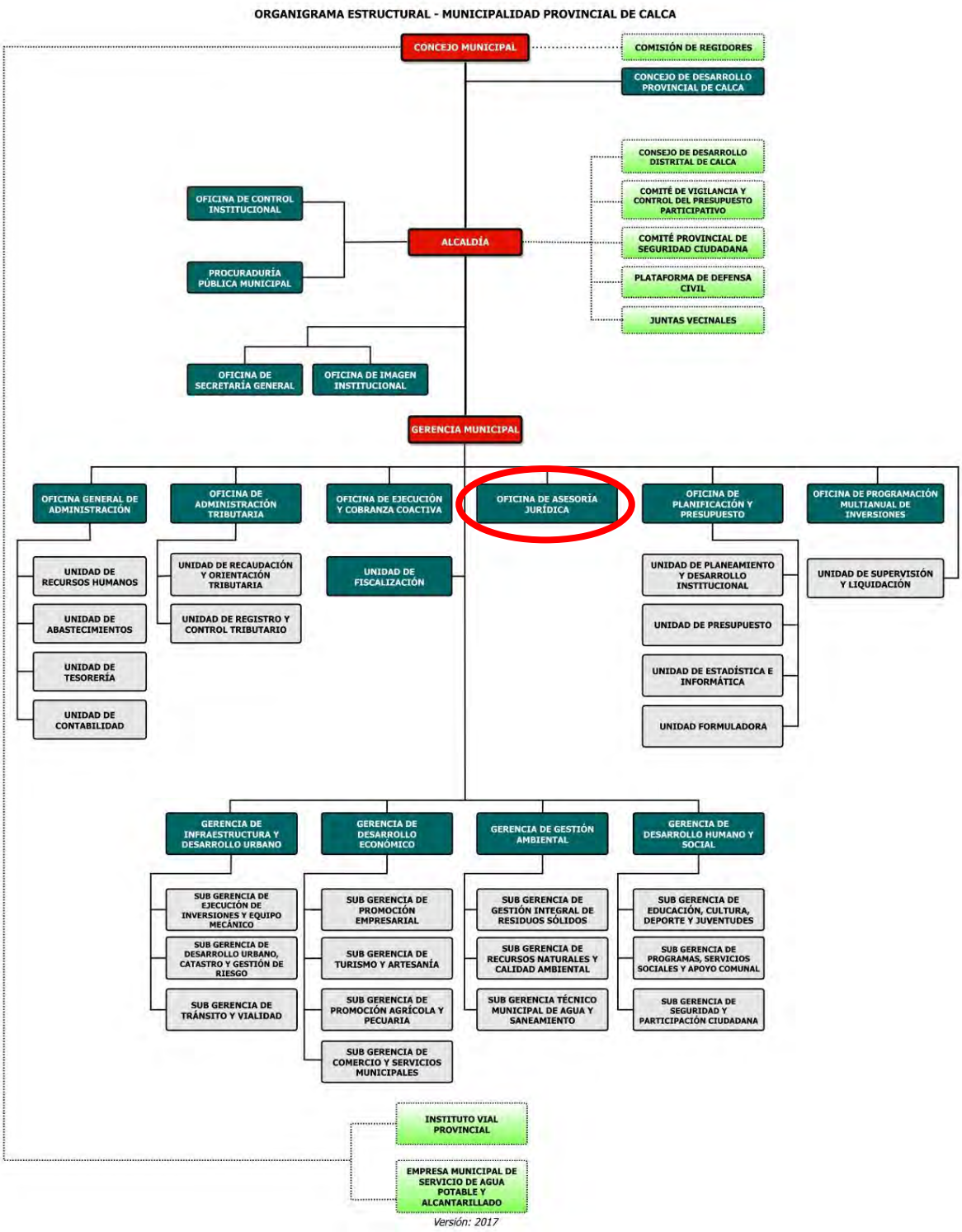
La Municipalidad Provincial de Calca comprometida con la mejora de las condiciones de vida de la población, a través de la promoción del Desarrollo Social, Económico, prestación de servicios de calidad, ejecutando obras priorizadas por la población de manera concertada y con una administración eficiente, eficaz y transparente.

2.7.2 Visión

La Municipalidad Provincial de Calca, es un Gobierno líder del Valle Sagrado de los Incas, promotora de la integración Provincial, con gestión eficiente, transparente y concertadora; priorizando los ejes temáticos de desarrollo del Plan de Desarrollo Concertado y respetando el Medio Ambiente, fortaleciendo la gestión de riesgos y la Mancomunidad Willcamayo.

2.7.3 Organigrama

Figura 1. Organigrama de la municipalidad provincial de Calca



Nota: Recuperado de MOF de la municipalidad provincial de Calca, 2024.

2.7.4 Oficina de Asesoría Jurídica

La Oficina de Asesoría Jurídica, es el órgano de asesoramiento encargado de dirigir y evaluar los asuntos de carácter legal de la Municipalidad, sobre la base de una adecuada interpretación de las normas legales (Calca, Manual de Organización y Funciones, 2024).

De las funciones específicas de la Oficina de Asesoría Jurídica:

- Asesorar a la Alta Dirección y demás órganos de la Municipalidad en asuntos jurídicos de carácter complejo.
- Elaborar, proponer y revisar proyectos de disposiciones legales que le solicite la Alta Dirección para una eficaz y eficiente Gestión Municipal, en concordancia con la legislación vigente.
- Asistir a las sesiones del Concejo Municipal con el fin de absolver las consultas de carácter legal.
- Emitir opiniones legales en los recursos impugnativos de apelación, que en segunda instancia administrativa deba resolver la Alta Dirección.
- Emitir opinión legal respecto a la elaboración y o actualización de los documentos de gestión institucional.
- Orientar hacia criterios de interpretación común, a los asesores jurídicos de la Institución, respecto a las opiniones legales que éstos elaboren.
- Informar a los diversos órganos de la municipalidad sobre las modificaciones legales que tenga implicancia en el desempeño de sus funciones.
- Participar en la formulación o revisión de contratos o convenios que encomiende la Alta Dirección.

- Cumplir con las demás funciones que le sean encomendadas por la Gerencia Municipal.

2.8 Generalidades de estructura de modelos

2.8.1 Tokenización

La tokenización es un proceso fundamental en el procesamiento del lenguaje natural (NLP) y los modelos de lenguaje grandes (LLMs), actuando como el puente entre el texto sin procesar y los datos comprensibles por máquinas. Consiste en dividir el texto en unidades más pequeñas, conocidas como tokens, que pueden ser palabras, subpalabras o caracteres, dependiendo del método elegido. Este proceso es crucial para permitir que los modelos interpreten y generen lenguaje humano de manera efectiva, como se detalla en el libro "Speech and Language Processing" de (Jurafsky & Martin, 2021)

2.8.2 Tokenización de Subpalabras

La *tokenización por subpalabras* es una técnica que segmenta palabras enteras en unidades más pequeñas—*subpalabras*—de manera que las palabras frecuentes se conservan intactas, mientras que las muy poco comunes se descomponen en fragmentos con significado. Esto permite que los modelos lingüísticos manejen un vocabulario compacto, mitiguen el problema de palabras fuera de vocabulario (*OOV*) y aún representen nuevos términos mediante combinaciones conocidas. (Jurafsky & Martin, 2021)

2.8.3 SFTTrainer

En el proceso de ajuste fino supervisado (Supervised Fine-Tuning, SFT), se entrena un modelo preentrenado utilizando pares de (instrucción, respuesta), de forma que éste aprenda a responder adecuadamente en un formato conversacional y especializarse en tareas específicas. Este método constituye un puente entre el conocimiento general del modelo y su aplicación efectiva en dominios concretos. (Zhou, Gopalakrishnan, Banerjee, & Narayan, 2023).

2.8.4 Input_ids

Los input_ids son secuencias de números enteros que representan tokens preprocesados por el tokenizer. Cada valor corresponde a la posición única de un token (palabra, subpalabra o símbolo) en el vocabulario del modelo. Estos identificadores constituyen el principal insumo que los modelos de lenguaje reciben para procesar texto. (Hugging Face, s.f.)

2.8.5 Attention_mask

La attention_mask es un mecanismo esencial usado en los modelos tipo transformer para indicar al modelo qué tokens debe considerar durante el cálculo de atención y a cuáles debe ignorar. Esta máscara se presenta como un tensor binario donde:

- El valor 1 señala que el token es relevante y el modelo debe atenderlo.
- El valor 0 indica que el token debe ser ignorado, como suelen ser los tokens de relleno (padding) o futuros tokens en tareas autorregresivas (causal masking)

Este enfoque asegura que el modelo no se vea influenciado por elementos irrelevantes, como espacios añadidos solo para normalizar la longitud de secuencias, y mantiene la coherencia en tareas donde solo debe utilizar información pasada. (Modular Inc., 2025).

2.8.6 Librerías – en general

2.8.6.1 Unsloth

Unsloth es una biblioteca de código abierto escrita en Python, diseñada para optimizar y acelerar el fine-tuning de modelos de lenguaje grande (LLMs). Gracias a derivaciones matemáticas manuales y kernels de GPU personalizados, Unsloth mejora significativamente la eficiencia del entrenamiento sin cambios de hardware, permitiendo procesos más rápidos y menos demandantes en memoria (Unsloth, 2024)

2.8.6.2 Torch

La librería Torch, conocida actualmente como PyTorch, es un marco de trabajo (framework) de código abierto diseñado para el desarrollo de modelos de aprendizaje profundo y cálculo numérico avanzado. Está basada en tensores multidimensionales y se caracteriza por el uso de grafos computacionales dinámicos (define-by-run), lo que permite construir, modificar y depurar redes neuronales en tiempo real. Además, incorpora un sistema de diferenciación automática (autograd) que facilita el cálculo de gradientes, optimizando así los procesos de entrenamiento de modelos. Desde el año 2022, PyTorch forma parte de la PyTorch Foundation, bajo la administración de la Linux Foundation, consolidándose como una de las bibliotecas más utilizadas en investigación académica y aplicaciones industriales de inteligencia artificial (Bergmann & Stryker, 2023).

2.8.6.3 Transformers

La librería Transformers, desarrollada por Hugging Face, es un marco de trabajo de código abierto que ofrece modelos preentrenados basados en la arquitectura Transformer, como BERT, GPT, T5, RoBERTa, DistilBERT, entre otros. Facilita la implementación de tareas de procesamiento de lenguaje natural (PLN), visión y audio mediante una API unificada y compatible con PyTorch y TensorFlow (Wolf, y otros, 2020).

2.8.6.4 Peft

PEFT es una librería de Hugging Face orientada al ajuste eficiente de parámetros en modelos de lenguaje de gran tamaño (LLMs). Su objetivo es reducir los recursos computacionales necesarios durante el fine-tuning, utilizando técnicas como LoRA (Low-Rank Adaptation), Prefix Tuning y Adapters, sin necesidad de ajustar todos los parámetros del modelo base (Mangrulkar, y otros, 2022).

2.8.6.5 Huggingface

Hugging Face es una empresa tecnológica y comunidad de código abierto dedicada al desarrollo de herramientas de inteligencia artificial, especialmente en el ámbito del procesamiento de lenguaje natural. Su ecosistema incluye la librería Transformers, el repositorio de modelos Model Hub, y bibliotecas asociadas como Datasets, PEFT, Accelerate y TRL, que permiten entrenar, compartir y desplegar modelos de manera colaborativa (Hugging Face, 2023).

2.8.6.6 Transformer Reinforcement Learning

La librería TRL (Transformer Reinforcement Learning) de Hugging Face está diseñada para el entrenamiento de modelos de lenguaje grandes (LLMs) con técnicas de aprendizaje por refuerzo, como Reinforcement Learning with Human Feedback (RLHF). Proporciona implementaciones de algoritmos como PPO (Proximal Policy Optimization) y permite adaptar modelos de lenguaje para alinearlos mejor con las preferencias humanas (von Werra, Scheurer, Wolf, Scao, & Raffel, 2020).

2.8.6.7 FastLenguajeModel

El FastLanguageModel es una clase desarrollada en la librería Unsloth que actúa como punto de entrada principal para la carga de modelos de lenguaje preentrenados con optimizaciones avanzadas de memoria y velocidad. Su diseño permite la integración con técnicas de ajuste fino eficiente como LoRA/PEFT, así como el uso de cuantización en 4 y 8 bits, lo que reduce significativamente los costos computacionales y facilita el uso de modelos de gran tamaño incluso en hardware limitado. Además, mantiene compatibilidad con la interfaz de Hugging Face Transformers, lo que asegura flexibilidad en el entrenamiento y la inferencia (Datacamp., 2024) (Deepwiki, 2024)

2.8.6.8 Tipo datos Dtype

El término dtype (del inglés data type) hace referencia al tipo de datos que contiene un tensor en librerías de aprendizaje profundo como PyTorch. Define el formato numérico de cada elemento del tensor —por ejemplo, si los valores son de punto flotante de 32 bits (float32), enteros de 64 bits (int64), booleanos, etc.— lo que afecta directamente la precisión aritmética, el uso de memoria y la eficiencia computacional en el entrenamiento e inferencia de modelos. (PyTorch, 2025)

Uso de dtype permite:

- Especificar el tipo de dato al crear tensores o al convertir entre tipos existentes.
- Controlar el balance entre precisión y rendimiento: tipos de datos de menor precisión (como float16) pueden reducir memoria y acelerar cómputos, aunque con mayor riesgo de error numérico.
- Asegurar compatibilidad ulterior en operaciones tensoriales, ya que muchos métodos requieren que los tensores tengan tipos de datos compatibles.

2.8.6.9 SafeTensor

Es un formato moderno para almacenar y cargar tensores de manera segura, eficiente y sin necesidad de deserialización arbitraria. Fue creado como alternativa a formatos basados en pickle, que presentan vulnerabilidades de seguridad. safetensors permite:

- Lectura y escritura más rápidas (zero-copy).
- Mayor seguridad, ya que no ejecuta código al cargar.
- Compatibilidad con librerías de modelos como Hugging Face Transformers.

De acuerdo con la documentación oficial de (Hugging Face, 2023) el formato está pensado para escalabilidad y despliegue de modelos en producción.

2.8.6.10 Lora Alpha

Alpha (α) es un parámetro de escala utilizado en LoRA que controla la magnitud de la contribución de las matrices de adaptación de bajo rango al modelo base. En la fórmula de actualización, después de determinar las matrices entrenables A y B de rango reducido, la actualización se introduce escalada por α/r , lo que permite modular cuánto aporta la parte adaptativa respecto al peso original. Un valor mayor de α implica una influencia mayor de los componentes adaptativos. (Lin, y otros, 2024)

2.8.6.11 Lora Dropout

LoRA Dropout es un mecanismo propuesto para regularizar el ajuste fino con LoRA, introduciendo un "apagado aleatorio" (dropout) en las activaciones de las matrices adaptativas de bajo rango durante el entrenamiento. Este método ayuda a controlar el sobreajuste (overfitting) al impedir que la adaptación memorice excesivamente los datos de entrenamiento. Científicos han demostrado que LoRA Dropout puede mejorar la generalización al introducir cierta esparsidad y variación en las actualizaciones. (Lin, y otros, 2024)

2.8.6.12 Eos_token

Es un token especial que indica el final de una secuencia de texto. Su función es crucial porque le permite al modelo reconocer el límite de un documento, oración o respuesta. Sin un token de fin de secuencia, el modelo podría generar texto de forma indefinida.

Según (Jurafsky & Martin, 2023) en Speech and Language Processing, los tokens especiales como <eos> y <bos> (beginning of sequence) son convenciones fundamentales en el preprocesamiento de datos para modelos de lenguaje.

2.8.6.13 Chunk token

Un chunk token se define como la unidad mínima de texto que integra un chunk o fragmento mayor en el Procesamiento de Lenguaje Natural (PLN). Estos tokens, al ser

agrupados en fragmentos, permiten que los modelos de lenguaje manejen textos extensos de manera más eficiente, superando las limitaciones de memoria y de ventana de contexto, y asegurando la coherencia semántica de cada división (Jaiswal & & Milios, 2023)

2.8.6.14 Chunk text

El chunk text hace referencia al propio fragmento de texto que resulta de dividir un documento mayor en secciones más pequeñas. Estos chunks suelen definirse bajo criterios de tamaño en palabras o tokens (ejemplo: fragmentos de 500 a 1000 palabras) para garantizar que la información sea manejable dentro de los límites de memoria y contexto del modelo. Esta técnica es ampliamente utilizada en aplicaciones de procesamiento de lenguaje natural (PLN) para mejorar la comprensión y mantener coherencia en textos largos (Lewis, y otros, 2020).

2.8.6.15 Chunking

El chunking es el proceso de dividir un texto o documento extenso en fragmentos más pequeños y manejables denominados chunks. Esta técnica busca optimizar el procesamiento y la recuperación de información al evitar que el modelo pierda contexto o sobrecargue su capacidad de análisis. En el ámbito de los grandes modelos de lenguaje (LLMs), el chunking es fundamental para tareas como la búsqueda semántica, la generación de resúmenes y la elaboración de respuestas a partir de documentos largos (Raffel, y otros, 2020; Lewis, y otros, 2020).

2.9 Fine Tuning

El Fine-Tuning es un proceso avanzado de transfer learning mediante el cual se toma un modelo de lenguaje previamente preentrenado sobre gran cantidad de datos generales y se ajusta utilizando un conjunto de datos específicos del dominio o tarea. El objetivo es adaptar sus capacidades para cumplir con precisión requisitos concretos, sin necesidad de partir de cero. Esto permite obtener modelos especializados, eficientes y contextualizados, ahorrando recursos

computacionales y mejorando su utilidad en escenarios reales. (Xu, Xie, Qin, Tao, & Wang, 2023)

2.9.1 Tipos de FineTuning

2.9.1.1 Fine-Tuning Completo (Full Fine-Tuning)

El Full Fine-Tuning consiste en reentrenar todos los parámetros de un modelo de lenguaje grande (LLM) utilizando un corpus de datos adaptado a la tarea objetivo. Este enfoque es considerado el método más directo y completo para adaptar modelos, ya que permite una especialización profunda; sin embargo, es también el más costoso en términos de cómputo, memoria y tiempo de entrenamiento, lo que lo hace menos eficiente frente a enfoques recientes como LoRA o PEFT (Zhang, Li, Zhao, & Chen, 2023).

2.9.1.2 Fine-Tuning Eficiente en Parámetros (PEFT – Parameter-Efficient Fine-Tuning)

El Fine-Tuning Eficiente en Parámetros (PEFT) es una estrategia de adaptación de modelos de lenguaje grandes (LLMs) que busca ajustar únicamente un pequeño subconjunto de parámetros del modelo preentrenado, en lugar de reentrenar todos sus pesos. Esto permite reducir significativamente el costo computacional, el consumo de memoria y los riesgos de sobreajuste, manteniendo un desempeño competitivo frente al Full Fine-Tuning. Entre las técnicas más utilizadas se incluyen LoRA (Low-Rank Adaptation), Prompt Tuning, y Adapters, las cuales permiten especializar el modelo en nuevas tareas sin perder su conocimiento general (Ding, y otros, 2023)

2.9.1.2.1 Subtipos principales:

- ***LoRA (Low-Rank Adaptation):***

LoRA (Adaptación de Bajo Rango) es un método diseñado para ajustar eficientemente modelos grandes mediante la introducción de matrices de bajo rango en las matrices de pesos del modelo. El objetivo principal de LoRA es reducir el coste computacional y el uso de

memoria durante el proceso de ajuste sin sacrificar la capacidad del modelo para aprender información específica de la tarea. Funciona añadiendo matrices de bajo rango entrenables a los pesos del modelo existente, en lugar de actualizar el conjunto completo de pesos. (Rathod, 2024)

- ***Adapters:***

Añaden módulos adicionales pequeños entre capas del modelo. El entrenamiento se limita a estos módulos, mientras el resto permanece fijo. (Hu, y otros, 2021)

- ***Prompt-Tuning / Prefix-Tuning:***

En lugar de modificar parámetros, se entrenan prompts blandos (vectores) que guían la generación del modelo. (Hu, y otros, 2021)

2.9.1.3 Fine-Tuning de Representación (Representation Fine-Tuning – ReFT)

El Fine-Tuning de Representación (ReFT) es una técnica de ajuste de modelos de lenguaje de gran escala (LLMs) que consiste en modificar las representaciones internas o vectores ocultos del modelo en lugar de actualizar directamente sus parámetros entrenables.
En este enfoque, el modelo base permanece congelado, y se entrenan únicamente pequeñas funciones de intervención que se insertan en determinadas capas para transformar las representaciones intermedias durante el proceso de inferencia.
ReFT permite lograr ajustes altamente eficientes y modulares, ya que "aprende a modificar activaciones internas específicas, en lugar de alterar la totalidad de los pesos del modelo", lo cual proporciona mayor interpretabilidad y un uso mucho más reducido de parámetros en comparación con métodos como LoRA o el fine-tuning completo. (Wu, Zou, & Rush, 2024)

2.9.1.4 Instruction Fine-Tuning (Supervised Fine-Tuning – SFT)

El Instruction Fine-Tuning, dentro de la categoría de Supervised Fine-Tuning (SFT), se refiere a la técnica mediante la cual un modelo de lenguaje preentrenado se ajusta empleando

un conjunto de datos estructurado como pares instrucción → respuesta. Según el LLM Engineer's Handbook, este enfoque permite que el modelo aprenda a entender y responder en un formato conversacional específico, fortaleciendo su rol como agente conversacional, y al mismo tiempo adaptar su conocimiento general a tareas o dominios especializados. (Iusztin & Labonne, 2024)

2.9.1.5 Fine-Tuning con Refuerzo a partir de Retroalimentación Humana (RLHF – Reinforcement Learning with Human Feedback)

El Fine-Tuning con Refuerzo a partir de Retroalimentación Humana (RLHF) constituye un enfoque híbrido que busca alinear los modelos de lenguaje con las preferencias humanas mediante la integración de retroalimentación directa de evaluadores humanos en el ciclo de entrenamiento. Este método combina un ajuste supervisado inicial con la construcción de un modelo de recompensa y la posterior optimización mediante algoritmos de aprendizaje por refuerzo como el PPO (Proximal Policy Optimization).

De acuerdo con (Ouyang, y otros, 2022)esta técnica permite que el modelo no solo responda de forma correcta desde un punto de vista gramatical o factual, sino que además lo haga de manera útil, segura y alineada con las expectativas humanas, representando un avance fundamental frente a los métodos tradicionales de ajuste fino.

2.9.1.6 Comparación General de Tipos de Fine-Tuning

Al comparar los distintos tipos de Fine-Tuning se pueden establecer criterios fundamentales que permiten evaluar su eficacia y aplicabilidad. Entre los más relevantes destacan: los parámetros ajustados durante el entrenamiento, el costo computacional requerido y el riesgo de que el modelo olvide el conocimiento general previamente adquirido (catastrophic forgetting).

Considerando estos aspectos, se presenta el siguiente cuadro comparativo que resume las características más importantes de cada enfoque de Fine-Tuning y sus principales ventajas.

Tabla 1 *Características más importantes de cada enfoque de Fine-Tuning*

Tipo	**Parámetros ajustados**	**Costo computacional**	**Riesgo de olvidar conocimiento general**	**Ejemplo de uso**
Full Fine-Tuning	100%	Muy alto	Alto	Especialización total en derecho municipal
PEFT (LoRA, Adapters, Prompt)	1% – 10%	Bajo – Medio	Bajo	Ajustar un modelo general para dictámenes
ReFT	<1%	Muy bajo	Muy bajo	Ajuste rápido en entornos limitados
Instruction Fine-Tuning (SFT)	Variable	Medio	Medio	Entrenar un modelo para seguir instrucciones
RLHF	Variable + política de refuerzo	Alto	Bajo	Ajustar modelo a preferencias humanas

Nota: Recuperado de LLM Engeneer's Handbook: Master the Art of Engineering Large Languaje Models from concept to production(p. 145), por Lusztin y M. Labonne, 2024.

Lo expuesto implica que, para el presente proyecto, se optará por implementar el enfoque de Fine-Tuning con LoRA (Low-Rank Adaptation), dado que representa una alternativa altamente eficiente en términos de parámetros y costo computacional. Este método permite especializar el modelo en el dominio legal sin necesidad de reentrenar todos sus parámetros, reduciendo así el riesgo de olvidar el conocimiento general y asegurando un mejor aprovechamiento de los recursos disponibles.

2.9.2 Overfitting

El sobreajuste (overfitting) ocurre cuando un modelo de aprendizaje automático aprende demasiado bien los datos de entrenamiento, capturando no solo los patrones relevantes, sino también el ruido y las particularidades irrelevantes. Esto conduce a un buen desempeño en el conjunto de entrenamiento, pero un rendimiento deficiente en datos nuevos o no vistos. Según Chollet (2021), "un modelo sobreajustado pierde capacidad de generalización al ajustarse en

exceso a los detalles de los datos de entrenamiento" (Deep Learning with Python, 2da ed., Manning).

2.9.3 Training Loss

La pérdida de entrenamiento (training loss) es una métrica que mide el error promedio del modelo sobre los datos utilizados durante el entrenamiento. Refleja qué tan bien el modelo está aprendiendo los patrones presentes en dicho conjunto. Como señalan Goodfellow, Bengio y Courville (2021), "la función de pérdida en el entrenamiento proporciona una estimación de la discrepancia entre las predicciones del modelo y los valores esperados, actuando como señal principal de ajuste de parámetros" (Deep Learning, MIT Press).

2.9.4 Validation Loss

La pérdida de validación (validation loss) mide el error del modelo al evaluarse en un conjunto de datos separado, que no se usó durante el entrenamiento. Sirve como indicador de la capacidad de generalización. Cuando la pérdida de validación comienza a aumentar mientras la pérdida de entrenamiento sigue disminuyendo, se considera un signo claro de sobreajuste. Según Brownlee (2022), "la pérdida de validación actúa como un barómetro de generalización, mostrando si el modelo está aprendiendo patrones útiles o simplemente memorizando" (Machine Learning Mastery).

2.10 Ingeniería de PROMPT

La ingeniería de prompt (Prompt Engineering) es una disciplina dentro de la Inteligencia Artificial que se encarga del diseño, formulación y optimización de instrucciones (prompts) dirigidas a modelos generativos, con el propósito de obtener respuestas precisas, coherentes y relevantes.

De acuerdo con (Encyclopaedia, 2024) la ingeniería de prompts es el proceso de diseñar entradas para modelos de inteligencia artificial generativa con el fin de producir respuestas útiles, precisas y relevantes.

En la misma línea, (TechTarget, 2025) define la ingeniería de prompts como una técnica de la inteligencia artificial que consiste en refinar las instrucciones dadas a modelos de lenguaje para optimizar la generación de contenido en sistemas de IA generativa.

Asimismo, desde un enfoque más técnico, diversas investigaciones académicas señalan que la ingeniería de prompts permite adaptar modelos preentrenados a nuevas tareas mediante instrucciones específicas, sin necesidad de modificar sus parámetros internos, lo que facilita su aplicación en múltiples dominios.

En este sentido, la ingeniería de prompts puede entenderse como una forma de programación en lenguaje natural, donde el usuario guía el comportamiento del modelo mediante instrucciones estructuradas, contexto y restricciones.

La ingeniería de prompts es fundamental en el uso de modelos de Inteligencia Artificial Generativa, ya que la calidad de las respuestas depende directamente de la calidad de las instrucciones proporcionadas.

Según (McKinsey & Company., 2024), diseñar adecuadamente los prompts permite obtener resultados óptimos, ya que mejores entradas generan mejores salidas en los modelos de IA generativa.

Asimismo, esta técnica permite:

- Mejorar la precisión y coherencia de las respuestas
- Reducir ambigüedades en la interpretación del modelo

- Adaptar el modelo a contextos específicos (como el jurídico parte de este trabajo de investigación)
- Optimizar el rendimiento sin necesidad de reentrenamiento

Adicionalmente, la literatura científica reciente destaca que la ingeniería de prompts se ha convertido en una habilidad clave para interactuar eficazmente con modelos de lenguaje, permitiendo resolver tareas complejas mediante estrategias como el uso de contexto, ejemplos y razonamiento guiado (Wei, y otros, 2022).

2.10.1 Clasificación de técnicas de prompting

2.10.1.1 Técnicas básicas de prompting

- **Zero-shot**

El Zero-shot prompting es una técnica de ingeniería de prompts que permite a los modelos de lenguaje ejecutar tareas sin haber recibido ejemplos específicos durante la interacción, basándose únicamente en la instrucción proporcionada por el usuario y en el conocimiento previamente adquirido durante su entrenamiento en grandes volúmenes de datos (Liu P. Y., 2023)Esta técnica se fundamenta en la capacidad de generalización de los modelos de lenguaje de gran escala, los cuales pueden inferir patrones y aplicar conocimientos previos a nuevas tareas sin necesidad de ajuste fino o entrenamiento adicional (Brown & Amodei, 2020) En este contexto, el Zero-shot prompting representa una forma eficiente de adaptar modelos preentrenados a distintos dominios, ya que permite obtener respuestas relevantes mediante el uso de instrucciones claras y bien estructuradas. No obstante, su efectividad depende en gran medida de la calidad y precisión del prompt, siendo más adecuado para tareas generales o de menor complejidad, en comparación con técnicas que incorporan ejemplos o razonamiento explícito (Kalyan & Rajendran, 2024)

- **Few-shot**

El Few-shot prompting es una técnica de ingeniería de prompts que consiste en proporcionar al modelo de lenguaje uno o varios ejemplos dentro del prompt con el objetivo de guiar su comportamiento y mejorar la calidad de las respuestas generadas (Brown & Amodei, 2020) A diferencia del enfoque zero-shot, esta técnica permite al modelo identificar patrones, estructuras y relaciones a partir de los ejemplos proporcionados, facilitando así la adaptación a tareas específicas sin necesidad de un proceso de entrenamiento adicional (Liu P. Y., 2023)En este sentido, el Few-shot prompting se basa en el principio de aprendizaje en contexto (*in-context learning*), mediante el cual los modelos de lenguaje de gran escala son capaces de inferir reglas implícitas a partir de un número reducido de ejemplos y aplicarlas a nuevas situaciones (Brown & Amodei, 2020) Asimismo, diversos estudios recientes destacan que esta técnica mejora significativamente la precisión, coherencia y consistencia de las respuestas, especialmente en tareas complejas o especializadas, como la redacción de textos técnicos o jurídicos (Kalyan & Rajendran, 2024) No obstante, su eficacia depende de la calidad, relevancia y representatividad de los ejemplos incluidos en el prompt, ya que ejemplos mal diseñados pueden inducir errores o sesgos en la salida generada (Liu P. Y., 2023)

- **Role prompting**

El Role prompting es una técnica de ingeniería de prompts que consiste en asignar un rol o identidad específica al modelo de lenguaje con el fin de orientar su comportamiento, estilo de respuesta y nivel de especialización en la tarea solicitada (Liu P. Y., 2023) Mediante esta técnica, se instruye al modelo para que actúe como un experto en un área determinada —por ejemplo, abogado, médico o ingeniero— lo que permite generar respuestas más contextualizadas, coherentes y alineadas con el dominio requerido. En este sentido, el Role prompting aprovecha la capacidad de los modelos de lenguaje para adaptar su discurso en

función del contexto proporcionado, mejorando la calidad de la salida generada (Kalyan & Rajendran, 2024) Asimismo, diversos estudios señalan que la asignación de roles contribuye a reducir ambigüedades en la interpretación de la tarea y a incrementar la precisión en entornos especializados.

2.10.1.2 Técnicas de razonamiento

- **Chain of Thought (CoT)**

El Chain of Thought (CoT) es una técnica de ingeniería de prompts que permite a los modelos de lenguaje generar respuestas mediante un proceso de razonamiento paso a paso, descomponiendo problemas complejos en una secuencia lógica de inferencias intermedias (Wei, y otros, 2022) Esta técnica se basa en inducir al modelo a explicar su razonamiento antes de proporcionar la respuesta final, lo que mejora significativamente la precisión en tareas que requieren análisis, como problemas matemáticos, lógicos o jurídicos. En este sentido, el CoT aprovecha la capacidad de los modelos de lenguaje de gran escala para simular procesos cognitivos estructurados, facilitando la interpretación y resolución de problemas complejos mediante la explicitación del razonamiento (Liu P. Y., 2023). Asimismo, diversos estudios han demostrado que el uso de CoT incrementa la coherencia y calidad de las respuestas, especialmente cuando se combina con otras técnicas como Few-shot prompting o Role prompting (Kalyan & Rajendran, 2024).

- **Variantes CoT (breve)**

Diversas investigaciones han propuesto variantes del Chain of Thought con el objetivo de mejorar su eficacia y adaptabilidad en distintos contextos. Entre las principales variantes se encuentra el Self-Consistency, que consiste en generar múltiples cadenas de razonamiento y seleccionar la respuesta más consistente, lo que permite reducir errores y aumentar la confiabilidad del resultado (Wang, Wei, & Schuurmans, 2022). Otra variante relevante es el

Least-to-Most prompting, que descompone un problema complejo en subtareas más simples que se resuelven progresivamente, facilitando el razonamiento estructurado en tareas de alta complejidad (Zhou, y otros, 2022) Asimismo, el Tree of Thoughts (ToT**)** amplía el enfoque de CoT al explorar múltiples caminos de razonamiento como un árbol de decisiones, permitiendo evaluar diferentes alternativas antes de llegar a una conclusión final (Yao, y otros, 2023)Estas variantes evidencian la evolución del CoT hacia enfoques más robustos que buscan mejorar la precisión, consistencia y capacidad de razonamiento de los modelos de lenguaje.

2.10.1.3 Técnicas de descomposición del problema

- **Plan and Solve**

El Plan and Solve (PaS) es una técnica de ingeniería de prompts que estructura el proceso de generación de respuestas en modelos de lenguaje mediante la formulación previa de un plan organizado que orienta la resolución de una tarea. Este enfoque se basa en la descomposición sistemática de problemas complejos en una secuencia lógica de pasos, permitiendo identificar los elementos relevantes, establecer un orden de ejecución y desarrollar respuestas de manera estructurada y coherente (Wang, y otros, 2023) La planificación previa actúa como un mecanismo de control que guía el proceso de generación, favoreciendo la claridad, consistencia y completitud de la información producida.

Asimismo, esta técnica permite optimizar la calidad de las salidas generadas al reducir la ambigüedad en la interpretación de la tarea, promoviendo una organización lógica del contenido y facilitando la trazabilidad del razonamiento seguido por el modelo. En este sentido, el Plan and Solve contribuye a mejorar la precisión, coherencia y estructuración de la información generada, especialmente en contextos donde se requiere un desarrollo ordenado y sistemático del contenido (Liu P. Y., 2023).

El uso de esta técnica presenta múltiples ventajas en la generación de contenido mediante modelos de lenguaje: Estructuración del contenido: Permite organizar la respuesta en pasos claros y definidos, mejora de la coherencia: Reduce respuestas desordenadas o incompletas, mayor precisión: Disminuye errores al seguir un plan lógico, trazabilidad del razonamiento: Facilita identificar cómo se construye la respuesta, puede aplicarse en distintos contextos, como redacción técnica, jurídica o análisis de datos.

Optimización del proceso: Reduce la necesidad de múltiples iteraciones de corrección.

- **Least-to-Most**

El Least-to-Most prompting es una técnica de ingeniería de prompts que permite a los modelos de lenguaje abordar problemas complejos mediante su descomposición en una secuencia progresiva de subtareas, iniciando con los elementos más simples y avanzando gradualmente hacia los más complejos. Este enfoque se basa en la idea de que la resolución escalonada facilita la comprensión del problema, permitiendo al modelo construir soluciones más precisas y coherentes a partir de la acumulación de resultados parciales (Zhou, y otros, 2022).

En este sentido, la técnica organiza el proceso de generación en etapas interdependientes, donde cada paso resuelto contribuye al siguiente, promoviendo una estructura lógica y acumulativa del conocimiento. Esta metodología favorece la claridad en la construcción de la respuesta y reduce la probabilidad de errores derivados de una comprensión incompleta del problema. Asimismo, el enfoque Least-to-Most permite mejorar la capacidad del modelo para manejar tareas que requieren múltiples niveles de razonamiento, al establecer una progresión ordenada desde lo simple hacia lo complejo (Liu P. Y., 2023)

2.10.1.4 Técnicas de verificación y control

- **Chain of Verification**

El Chain of Verification (CoV) es una técnica de ingeniería de prompts orientada a mejorar la confiabilidad y precisión de las respuestas generadas por modelos de lenguaje mediante la incorporación de un proceso estructurado de verificación posterior a la generación inicial. Este enfoque se basa en la idea de que el modelo no solo debe producir una respuesta, sino también evaluar su consistencia, detectar posibles errores y realizar correcciones cuando sea necesario, a través de una secuencia organizada de validación (Kalyan & Rajendran, 2024).

En este sentido, el CoV introduce una fase adicional en el proceso de generación, en la cual el modelo revisa críticamente la información producida, contrastándola con el contexto proporcionado y verificando la coherencia lógica, la ausencia de contradicciones y la adecuación al problema planteado. Esta técnica permite mejorar la calidad de las respuestas al reducir la presencia de errores, omisiones o inconsistencias, promoviendo un enfoque más riguroso y controlado en la generación de contenido (Liu P. Y., 2023).

Asimismo, el Chain of Verification contribuye a fortalecer la trazabilidad del proceso de generación, ya que permite identificar y justificar las correcciones realizadas, lo que resulta especialmente relevante en contextos donde se requiere un alto nivel de precisión y confiabilidad en la información generada. De esta manera, esta técnica favorece la producción de respuestas más sólidas, verificadas y alineadas con los criterios de calidad exigidos en entornos técnicos y especializados.

El uso de esta técnica presenta ventajas relevantes: Detecta y corrige inconsistencias en la respuesta, mayor confiabilidad generando resultados más precisos y verificables, introduce una fase de revisión sistemática, evita contradicciones internas, permite justificar y explicar las correcciones y aumenta el nivel técnico de las respuestas.

- **Self-refine**

El Self-Refine es una técnica de ingeniería de prompts que permite a los modelos de lenguaje mejorar progresivamente la calidad de sus respuestas mediante un proceso iterativo de generación, evaluación y refinamiento. Este enfoque se basa en la capacidad del modelo para autoevaluar sus propias salidas, identificar deficiencias o áreas de mejora y generar versiones revisadas que incrementen la coherencia, precisión y calidad del contenido producido (Madaan et al., 2023).

En este sentido, el proceso de Self-Refine se estructura en múltiples ciclos donde el modelo, en primer lugar, produce una respuesta inicial; posteriormente, realiza una evaluación crítica de dicha respuesta, identificando errores, ambigüedades o inconsistencias; y finalmente, genera una versión mejorada incorporando las correcciones necesarias. Este mecanismo permite simular un proceso de revisión similar al realizado por un experto humano, promoviendo una mejora continua del contenido generado sin necesidad de intervención externa (Madaan, y otros, 2023).

Asimismo, esta técnica contribuye significativamente a elevar la calidad de las respuestas en tareas que requieren precisión y claridad, ya que introduce un proceso de control interno que reduce errores y optimiza la estructura del contenido. De esta manera, el Self-Refine favorece la generación de resultados más elaborados, consistentes y alineados con los criterios de calidad exigidos en contextos técnicos y especializados (Liu P. Y., 2023).

2.10.1.5 Técnicas avanzadas

Exis Existen diversas técnicas avanzadas en el uso de modelos de lenguaje que permiten mejorar la calidad, coherencia y precisión de las respuestas generadas. Entre las más destacadas se encuentran:

- **Tree of Thoughts (Árbol de Pensamientos):** Técnica que permite al modelo explorar múltiples rutas de razonamiento antes de llegar a una respuesta final, evaluando diferentes alternativas de solución de manera estructurada.
- **ReAct(Reasoning Acting):** Combina el razonamiento con la ejecución de acciones, permitiendo al modelo interactuar con herramientas externas mientras desarrolla su proceso lógico para resolver problemas.
- **Meta prompting:** Consiste en diseñar instrucciones más complejas y estructuradas que guían al modelo sobre cómo debe razonar o responder, mejorando la calidad del resultado.
- **Self-consistency(Autoconsistencia):** Técnica que genera múltiples respuestas para una misma consulta y selecciona la más coherente o frecuente, aumentando la precisión del resultado final.

Asimismo, cabe señalar que existen muchas otras técnicas avanzadas en constante desarrollo, las cuales buscan optimizar el rendimiento de los modelos de inteligencia artificial en tareas específicas, especialmente en contextos complejos como el análisis y redacción de documentos legales.

Capítulo III: Marco metodológico

3.1 Tipo de la Investigación

La presente investigación es de tipo aplicada (Hernández Sampieri, 2018), debido a que se orienta a la solución de un problema concreto identificado en la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca. En este sentido, se desarrolla y evalúa una solución tecnológica basada en inteligencia artificial generativa, mediante modelos de lenguaje a gran escala (LLM), con el propósito de automatizar el proceso de emisión de opiniones legales. Este tipo de investigación busca la aplicación práctica del conocimiento, generando una propuesta funcional que contribuya a mejorar la eficiencia operativa en un contexto institucional real.

3.2 Nivel de Investigación

La investigación presenta un alcance descriptivo–explicativo con un diseño cuasi experimental.

Es descriptiva, en la medida en que caracteriza el proceso actual de emisión de opiniones legales, identificando sus principales limitaciones relacionadas con el procesamiento de expedientes administrativos, los tiempos de elaboración, la carga operativa y calidad de elaboración.

Asimismo, es explicativa, debido a que analiza el efecto de la implementación de una solución basada en inteligencia artificial generativa sobre el proceso de elaboración de opiniones legales. En ese sentido, se evalúa cómo la incorporación de un prototipo funcional basado en Modelos de Lenguaje a Gran Escala (LLM) influye en indicadores como la eficiencia del procesamiento de expedientes administrativos, la coherencia jurídica, la adecuación normativa, la fundamentación argumentativa, la claridad de la redacción y el tiempo de generación de las opiniones legales.

Adicionalmente, la investigación adopta un diseño cuasi experimental, debido a que se desarrolla e implementa un prototipo funcional basado en inteligencia artificial generativa, cuya eficiencia y calidad fueron evaluadas por especialistas jurídicos, permitiendo analizar los efectos de la intervención tecnológica en el proceso de elaboración de pre-dictámenes jurídicos.

3.3 Diseño metodológico de la investigación

La presente investigación adopta un diseño metodológico de tipo cuasi experimental, estructurado en función de los objetivos específicos. El procedimiento metodológico comprende cuatro etapas secuenciales: (i) construcción del dataset jurídico, (ii) comparación de modelos de lenguaje, (iii) desarrollo del prototipo funcional y (iv) evaluación de la calidad de las opiniones legales generadas.

Figura 2. *Diseño metodológico*



Nota: Imagen generada mediante inteligencia artificial y posteriormente editada en Canva (elaboración propia, 2026).

3.3.1 Construcción de dataset jurídico

En la primera etapa, se construyó un dataset a partir de 453 informes de opinión legal correspondientes al periodo 2022–2024, con la finalidad de entrenar, ajustar y evaluar el

desempeño de los modelos de lenguaje a gran escala utilizados en el trabajo de investigación. Para los modelos Llama y Mistral, el dataset fue estructurado en formato JSON, organizado en las categorías instruction, input y output. Por otro lado, para el modelo Claude se empleó el mismo conjunto de datos en formato de texto plano (TXT), adaptado a sus requerimientos de procesamiento.

3.3.2 Métricas de evaluación

La selección de las métricas de evaluación se fundamentó en la necesidad de comparar el desempeño de los modelos de lenguaje desde diferentes perspectivas. Se empleó BLEU, propuesta por (Papineni, 2002), para medir la coincidencia de secuencias de palabras entre los textos generados y los documentos de referencia. Asimismo, se utilizaron las métricas ROUGE-1, ROUGE-2 y ROUGE-L, propuestas por (Lin C.-Y. , 2004), con la finalidad de analizar la coincidencia de palabras, secuencias de dos palabras y subsecuencias comunes, respectivamente. Estas métricas permitieron establecer una comparación basada en las características textuales de las opiniones legales generadas.

De manera complementaria, se utilizó BERTScore, propuesto por (Artzi, 2020), que calcula la similitud entre los textos mediante representaciones contextuales de sus palabras. Su incorporación permitió complementar la evaluación basada en coincidencias léxicas con una medida de similitud semántica. Adicionalmente, se consideraron el tiempo de generación y la cantidad de tokens utilizados como indicadores operativos de eficiencia, debido a su relación con el tiempo de respuesta y el consumo de recursos durante la generación. En conjunto, estas métricas permitieron comparar cuantitativamente los modelos Llama, Mistral y Claude. No obstante, debido a que las métricas automáticas no verifican directamente la validez de las normas ni la corrección del razonamiento jurídico, sus resultados deben interpretarse como

indicadores de desempeño textual y computacional, y no como una medida absoluta de calidad jurídica.

3.3.3 Comparación de modelos de lenguaje

En la segunda etapa se comparó el desempeño de tres Modelos de Lenguaje a Gran Escala (LLM): Claude, Llama 3.1 y Mistral, con el propósito de identificar la alternativa más adecuada para la generación automatizada de opiniones legales en el contexto de la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca.

Para garantizar una evaluación objetiva del desempeño de los modelos, el dataset jurídico conformado por 453 expedientes administrativos fue dividido mediante una partición aleatoria reproducible (random_state = 42), destinando el 80 % (362 expedientes) al entrenamiento y el 20 % (91 expedientes) a la validación. Esta proporción constituye una práctica ampliamente utilizada en aprendizaje automático, ya que permite disponer de un número suficiente de ejemplos para el aprendizaje del modelo, preservando al mismo tiempo un conjunto independiente para evaluar su capacidad de generalización sobre datos no utilizados durante el entrenamiento (Amazon.com., 2026). La comparación entre los modelos se realizó utilizando exclusivamente los expedientes del conjunto de validación, garantizando que todos fueran evaluados sobre los mismos casos y bajo las mismas condiciones experimentales.

Para cada expediente se empleó la misma información de entrada, conformada por los antecedentes administrativos y jurídicos, así como una estructura uniforme de instrucciones (prompt), con la finalidad de asegurar que las diferencias observadas en los resultados dependieran únicamente del comportamiento de cada modelo y no de variaciones en los datos de entrada.

En el caso de los modelos Llama 3.1 y Mistral, se utilizaron versiones ajustadas mediante la técnica de fine-tuning, empleando el conjunto de entrenamiento previamente definido. Por su parte, Claude fue evaluado como modelo propietario sin ajuste de parámetros, utilizando técnicas avanzadas de ingeniería de prompts, específicamente Plan-and-Solve Prompting y Chain of Verification. Esta diferencia metodológica responde a las características propias de cada tecnología, ya que los modelos de código abierto permiten realizar procesos de ajuste supervisado mediante fine-tuning, mientras que los modelos propietarios restringen el acceso a sus parámetros internos y su adaptación se realiza principalmente mediante estrategias de ingeniería de prompts. En consecuencia, la comparación no se orientó únicamente a evaluar modelos de lenguaje, sino también las estrategias de adaptación técnicamente viables para cada uno de ellos.

La evaluación comparativa se realizó mediante métricas cuantitativas y cualitativas. Como métricas cuantitativas se emplearon BLEU, ROUGE, Exact Match, BERTScore, tiempo de inferencia y cantidad de tokens generados, comparando las opiniones legales generadas por cada modelo con las opiniones legales elaboradas por la Oficina de Asesoría Jurídica, consideradas como referencia (ground truth). Estas métricas permitieron medir la similitud léxica, la cobertura del contenido, la coincidencia exacta, la similitud semántica y la eficiencia computacional de cada modelo.

Complementariamente, se realizó una evaluación cualitativa considerando criterios relacionados con la integridad de la estructura del documento, la coherencia jurídica, la adecuación normativa, la precisión en la interpretación de los antecedentes y la claridad de la redacción jurídica, permitiendo valorar aspectos propios del razonamiento jurídico que no pueden ser medidos únicamente mediante métricas automáticas.

Los resultados obtenidos permitieron identificar el modelo con mejor equilibrio entre calidad de generación, precisión jurídica y eficiencia computacional, el cual fue seleccionado para su implementación en el prototipo funcional desarrollado en la siguiente etapa de la investigación. Es importante precisar que esta fase correspondió a una evaluación técnica comparativa de los modelos de lenguaje, cuyo propósito fue seleccionar la alternativa tecnológica más adecuada. En consecuencia, esta evaluación no constituyó la validación del sistema propuesto, la cual fue desarrollada posteriormente mediante una evaluación funcional realizada por seis especialistas jurídicos de la Oficina de Asesoría Jurídica.

3.3.4 Desarrollo del prototipo funcional

En la tercera etapa, se desarrolló un prototipo funcional basado en el modelo seleccionado, orientado a la automatización del procesamiento de expedientes administrativos y la generación de opiniones legales.

El sistema fue diseñado bajo un flujo secuencial que comprende:

- Carga de documentos en formato PDF
- Procesamiento mediante OCR
- Extracción de antecedentes relevantes
- Generación automatizada de opiniones legales mediante el modelo LLM

Asimismo, se incorporaron técnicas de ingeniería de prompts para mejorar la calidad de los resultados generados.

El prototipo fue implementado como una aplicación web, permitiendo automatizar el flujo de procesamiento documental desde la carga del expediente hasta la generación de la opinión legal.

3.3.5 Evaluación de la calidad de las opiniones legales generadas

En la cuarta etapa se evaluó el desempeño del prototipo funcional desarrollado mediante la valoración de la calidad jurídica de las opiniones legales generadas, así como de su usabilidad e impacto organizacional. Para ello, se aplicó una ficha de evaluación estructurada a seis (06) abogados especialistas jurídico pertenecientes a la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca, quienes participaron como evaluadores expertos.

El instrumento de evaluación fue diseñado para valorar el desempeño del sistema desde tres dimensiones: calidad técnica y coherencia jurídica, usabilidad del sistema e impacto organizacional. Para efectos del presente objetivo específico, referido a la evaluación de la calidad de las opiniones legales generadas, se consideraron los siguientes criterios:

- Coherencia jurídica del análisis.
- Adecuación normativa.
- Precisión en la interpretación de antecedentes.
- Fundamentación argumentativa.
- Estructura formal y redacción técnica.

Cada criterio fue valorado mediante una escala Likert de cinco niveles.

Tabla 2 *Escala Likert de cinco niveles*

Puntaje	Interpretación
1	Muy deficiente
2	Deficiente
3	Regular
4	Bueno
5	Excelente

Nota: Adaptado de diseño del formato de escala tipo Likert: un estado de la cuestión por Metas, A.2018.

Cada uno de los seis especialistas evaluó de manera independiente las opiniones legales generadas por el prototipo a partir de expedientes administrativos seleccionados para la investigación, asignando una calificación a cada criterio de evaluación.

Posteriormente, los puntajes individuales fueron consolidados mediante el cálculo del promedio aritmético por criterio y del promedio general del instrumento, permitiendo obtener una valoración cuantitativa del desempeño del sistema respecto a la calidad de las opiniones legales generadas. Los resultados obtenidos fueron representados mediante tablas comparativas para facilitar su análisis e interpretación.

Previamente a su aplicación, la ficha de evaluación fue sometida a un proceso de validación de contenido mediante juicio de expertos, participando un abogado especialista en Gestión Pública y un ingeniero de sistemas con experiencia en desarrollo de soluciones tecnológicas basadas en inteligencia artificial. Ambos revisaron la pertinencia, claridad, coherencia y suficiencia de los criterios establecidos, efectuándose los ajustes correspondientes antes de su aplicación a los especialistas jurídicos participantes.

No se aplicaron pruebas estadísticas inferenciales debido a que la evaluación correspondió a un juicio de expertos con una muestra no probabilística de tipo intencional de seis especialistas; en consecuencia, el análisis de los resultados se realizó mediante estadística descriptiva, utilizando frecuencias, promedios y tablas comparativas.

Capítulo IV: Desarrollo experimental del modelo IA

4.1 Selección de modelos LLM

La selección de los modelos de lenguaje se realizó considerando criterios de desempeño, accesibilidad, eficiencia computacional y aplicabilidad en tareas de procesamiento de lenguaje natural en el ámbito jurídico.

En primer lugar, se seleccionó el modelo LLaMA 3.1 (8B), desarrollado para tareas avanzadas de generación de texto, debido a su capacidad de adaptación mediante técnicas de fine-tuning y su eficiencia en entornos con recursos computacionales limitados. En segundo lugar, se empleó el modelo Mistral v0.3 (7B), el cual destaca por su arquitectura optimizada y su rendimiento competitivo en tareas de comprensión y generación de lenguaje natural.

Finalmente, se incluyó el modelo Claude Haiku, caracterizado por su alto desempeño en tareas de razonamiento y generación textual, el cual fue utilizado bajo un enfoque de inferencia directa mediante ingeniería de prompts. La inclusión de este modelo permite evaluar el desempeño de soluciones sin entrenamiento adicional frente a modelos ajustados específicamente para el dominio jurídico.

4.2 Construcción de dataset jurídico

4.2.1 Dataset para Llama y Mistral

Se diseñó un dataset estructurado en formato de pares entrada–salida (input–output), en el cual la entrada corresponde a la descripción del caso legal extraída de los expedientes administrativos, mientras que la salida consiste en la opinión legal elaborada por la Oficina de Asesoría Jurídica.

El dataset fue construido a partir de los 453 documentos recopilados correspondientes a informes de opinión legal emitidos entre los años 2022 y 2024, los cuales fueron procesados y organizados para su utilización en el entrenamiento y ajuste de los modelos Llama y Mistral.

Asimismo, se realizó un proceso de estructuración de la información, identificando componentes clave como antecedentes, análisis jurídico y conclusiones, con el propósito de mantener la coherencia técnica y el formato característico de las opiniones legales institucionales.

Este enfoque permitió disponer de un conjunto de datos representativo y alineado al contexto jurídico-administrativo, facilitando la generación de resultados consistentes durante el proceso de evaluación de los modelos.

4.2.2 Dataset para Claude

Para el modelo Claude se adoptó un enfoque basado en *prompt engineering*, en lugar del uso de un dataset para entrenamiento o ajuste fino del modelo.

En este caso, la información proveniente de los expedientes administrativos —derivada del conjunto de **y** fue incorporada directamente en prompts estructurados, los cuales incluyen la descripción del caso, antecedentes relevantes y lineamientos específicos para la generación de la opinión legal.

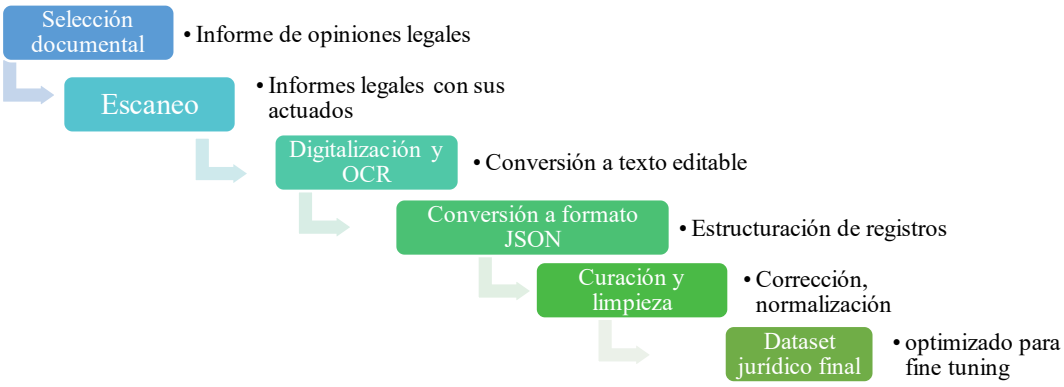
Este enfoque permite aprovechar la capacidad del modelo para comprender y procesar instrucciones complejas en lenguaje natural, sin requerir un proceso previo de entrenamiento adicional. Asimismo, se diseñaron prompts utilizando técnicas como *Plan and Solve* y *Chain of Verification*, con el propósito de mejorar la coherencia jurídica, la precisión normativa y la consistencia de los resultados generados.

De esta manera, el modelo opera sobre la información contextual proporcionada en cada ejecución, garantizando adaptabilidad a distintos casos sin necesidad de construir un dataset de entrenamiento específico, lo cual resulta adecuado para entornos donde se requiere flexibilidad y rapidez en la generación de documentos legales.

4.3 Preprocesamiento de datos

El proceso de preprocesamiento de datos tuvo como finalidad garantizar la calidad y consistencia del dataset utilizado en el entrenamiento de los modelos.

Figura 3 *Proceso de construcción dataset jurídico. Elaboración propia*



A continuación, se describen las etapas seguidas en este procedimiento:

4.3.1 Selección documental

Se realizó la recopilación de un total de 453 documentos correspondientes a informes de opinión legal emitidos por la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca durante el periodo comprendido entre los años 2022 y 2024.

El proceso de selección documental se efectuó mediante un muestreo intencional, considerando como criterio principal la inclusión de documentos representativos de la producción jurídica institucional. En este sentido, se priorizaron aquellos informes que reflejan la estructura formal, el lenguaje técnico-jurídico y los patrones de argumentación utilizados en la elaboración de opiniones legales.

Asimismo, la selección buscó garantizar la diversidad de casos en términos de complejidad administrativa y normativa, con el propósito de contar con un conjunto de datos adecuado para la evaluación del desempeño del sistema propuesto.

4.3.2 Escaneo de opiniones legales

Se escanearon las opiniones legales con sus respectivos actuados según la selección realizada. El formato que se utilizó para los archivos fue en Portable Document Format (PDF).

4.3.3 Digitalización y OCR

Los documentos en formato físico fueron digitalizados y, se aplicó Reconocimiento Óptico de Caracteres (OCR) para convertirlos en archivos editables. Esto permitió unificar el corpus en un formato digital homogéneo. Además en esta fase se aplicaron los procesos de corrección de errores ortográficos, combinación de caracteres y símbolos en mecanismos de reconocimiento, eliminación de encabezados de cada página entre otras.

4.3.4 Conversión a formato JSON

Debido a que LLaMA 3.1 requiere datos estructurados para el ajuste fino, se diseñó un esquema en formato **JSON**. Cada registro del dataset fue organizado con campos principales como:

```
[
  {
    "input": " ",
    "output": "Desarrollo de informe de opinión legal, estructurado de la siguiente de la siguiente manera: Antecedentes, Base legal, Análisis, Conclusiones y Recomendaciones"
    "instrucción": "Asunto de la opinión legal",
  }
]
```

Este formato asegura compatibilidad con las librerías de entrenamiento y facilita procesos de tokenización y batch processing en los modelos Llama y Mistral.

4.3.5 Curación y limpieza de datos

Con el fin de garantizar la calidad del dataset, se realizaron los siguientes pasos:

- Corrección ortográfica y de formato.
- Eliminación de documentos duplicados.
- Normalización terminológica jurídica.
- Filtrado de documentos irrelevantes para el dominio legal.

4.3.6 Resultado

Como resultado del proceso de recopilación, depuración y estructuración de la información, se obtuvo un dataset jurídico especializado, compuesto por un total de 453 documentos de opiniones legales emitidas por la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca, correspondientes al periodo 2022–2024.

- Este dataset fue transformado a un formato estructurado basado en esquemas de *instruction tuning* (instruction–input–output), donde:
- Instruction (instrucción): define la tarea a realizar, asociada al asunto o requerimiento jurídico del expediente.
- Input (entrada): contiene la descripción del caso, antecedentes administrativos y elementos relevantes para el análisis legal.
- Output (salida): corresponde a la opinión legal emitida, redactada conforme a los criterios técnicos y normativos institucionales.

Esta estructuración permitió establecer una correspondencia semántica y jurídica coherente entre los elementos del dataset, facilitando el proceso de aprendizaje supervisado durante el ajuste de los modelos y garantizando la generación de resultados alineados con el contexto jurídico institucional.

```
[
  {
    "output":"Mediante el presente me dirijo a Ud. en atención al Proveído dispuesto por el Despacho de Alcaldía, así como del Informe remitido por el Gerente de Desarrollo Humano y Social, sobre pronunciamiento legal del Reconocimiento del Comité del Programa del Vaso de Leche de la Municipalidad Provincial de Calca.\nI. ANTECEDENTES:\n1.1.Se tiene el Informe N° 059-2022/GDHS/JFJR/MPC de fecha 01 de Marzo del 2022 del Gerente de Desarrollo Humano y Social de esa fecha Abog. Julio Fernando Jaén Rodríguez, por el cual se solicita el RECONOCIMIENTO DEL COMITÉ DEL PROGRAMADEL VASO DE LECHE - 2022 de la Municipalidad Provincial de Calca.\nII. BASE LEGAL:\n• Constitución Política del Estado. • Ley N° 27972, Ley Orgánica de Municipalidades • Ley N° 27444, Ley General del Procedimiento Administrativo. • Ley N° 29662, Ley que crea el Sistema Nacional de Gestión de Riesgo de Desastres.\nIII. ANÁLISIS:\n3.1.Que, conforme a las facultades otorgadas a las Municipalidades en el artículo 194° de la Constitución Política del Perú, modificada por la Ley de Reforma Constitucional N° 28607, en Concordancia con lo establecido en el Artículo II del Título Preliminar de la Ley Orgánica de Municipalidades N° 27972, establece que los Gobiernos Locales gozan de Autonomía política, económica y administrativa en los asuntos de su competencia, radicando esta autonomía en la facultad de ejercer actos de gobierno administrativo y de administración, con sujeción al ordenamiento jurídico. 3.2. - Que, estando a lo establecido en el Artículo IV. Principios del procedimiento administrativo que preceptúa al Principio de legalidad: Las autoridades administrativas deben actuar con respeto a la Constitución, la ley y al derecho, dentro de las facultades que le estén atribuidas y de acuerdo con los fines para los que les fueron conferidas. 3- .3. Que, el artículo VIII del Título Preliminar de la Ley Orgánica de Municipalidades - Ley N° 27972, establece que, los gobiernos locales están sujetos a las Leyes y\ndisposiciones, que de manera general y de conformidad con la Constitución Política, regulan las actividades y funcionamiento del sector público, así como a las normas técnicas referidas a los servicios y bienes públicos, y a los sistemas administrativos de1 Estado que, por su naturaleza son de observancia y cumplimiento obligatorio; y, las competencias y funciones específicas municipales se cumplen en armonía con las políticas y planes nacionales, regionales y locales de desarrollo;; 3.4. - Que, la Ley Orgánica de Municipalidades en su CAPÍTULO II LAS NORMAS MUNICIPALES Y LOS PROCEDIMIENTOS ADMINISTRATIVOS SUBCAPÍTULO I LAS\nNORMAS MUNICIPALES Artículo 38.- ORDENAMIENTO JURÍDICO MUNICIPAL: El ordenamiento jurídico de las municipalidades está constituido por las normas emitidas por los órganos de gobierno y administración municipal, de acuerdo al ordenamiento jurídico nacional. Las normas y disposiciones municipales se rigen por los principios de exclusividad, territorialidad, legalidad y simplificación administrativa, sin perjuicio de la vigencia de otros principios generales del derecho administrativo. Ninguna autoridad puede avocarse a conocer o normar las materias que la presente ley orgánica establece como competencia exclusiva de las municipalidades. (...), Artículo 43.- RESOLUCIONES DE ALCALDÍA Las resoluciones de alcaldía aprueban y resuelven los asuntos de carácter administrativo;\n3.5. - Que, por otro lado la Ley orgánica de municipalidades en el Artículo 84.- PROGRAMAS SOCIALES, DEFENSA Y PROMOCIÓN DE DERECHOS: Las municipalidades, en materia de programas sociales, de defensa y promoción de derechos, ejercen las siguientes funciones: 1. Funciones específicas exclusivas de las municipalidades provinciales: 1.1. Planificar y promover el desarrollo social en su circunscripción en armonía con las políticas y planes nacionales y regionales, de manera concertada con las municipalidades distritales de su jurisdicción. 1.2. Establecer canales de concertación entre las instituciones que trabajan en defensa de derechos de niños y adolescentes, mujeres, discapacitados y adultos mayores. Así como de los derechos humanos en general,
```

manteniendo un registro actualizado. 1.3. Regular las acciones de las Defensorías Municipales de los Niños y Adolescentes, DEMUNA, adecuando las normas nacionales a la realidad local. 1.4. Ejecutar el Programa del Vaso de Leche y demás programas de apoyo alimentario con participación de la población y en concordancia con la legislación sobre la materia, cuando la municipalidad distrital no pueda asumir dicha función, (...); 3.6.- Que, el Artículo I o en concordancia con lo dispuesto en el artículo 2o, numeral 16, y en el artículo 30 ° de la Constitución Política, declaran de de prioritario interés nacional la labor que realizan los Clubes de Madres, Comités de Vaso de Leche, Comedores Populares Autogestionarios, Cocinas Familiares, Centros Familiares, Centros Materno-Infantiles y demás organizaciones sociales de base, en lo referido al servicio de apoya alimentario que brindan a las familias de menores recursos, así mismo de acuerdo al Artículo 4o de su Reglamento aprobado mediante Decreto Supremo N° 041-201 l-PCM,el Estado reconoce la existencia legal de las organizaciones sociales de base y les otorga personería jurídica privada de acuerdo al procedimiento descrito en el citado reglamento;\n3- .7. Que, el Artículo 2o de la Ley N° 27470, Ley que establece normas complementarias para la ejecución del Programa del Vaso de Leche, refiere (...) en cada Municipalidad Provincial, Distrital y Delegadas ubicadas en su jurisdicción, se conforma un Comité de Administración del Programa del Vaso de Leche, el mismo que es aprobado mediante\nResolución de Alcaldía con Acuerdo previo del Concejo Municipal, el cual está integrado por el Alcalde, un funcionario municipal, un representante de salud y tres representantes del Programa del Vaso de Leche, elegidas democráticamente por sus bases, de acuerdo a los estatutos de su organización y adicionalmente deberá estar conformado por un representante de la Asociación de Productores Agropecuarios de la región o zona, cuya representatividad será debidamente acreditada por el Ministerio de Agricultura (...), 2.2. las municipalidades como responsables de la ejecución del Programa del Vaso de Leche, en coordinación con la organización del vaso de leche, organizan programas, coordinan y ejecutan la implementación de dicho programa en sus fases de selección de beneficiarios, programación, distribución, supervisión y evaluación, 2.3. Los Gobiernos Locales podrán celebrar entre si Convenios para adquirir en forma conjunta los recursos inherentes al Programa del Vaso de Leche, a efecto de abaratar sus costos y/o mejorar la calidad de los mismos; 3- .8. Que, mediante Informe N° 059-2022/GDHS/JFJR/MPC de fecha 01 de marzo del presente año, el Gerente de Desarrollo Humano y Social, eleva el Informe N° 019-2022- PVL/MPC-GDS emitido por la Responsable del Programa del Vaso de Leche Tec. Nancy Callapiña Callasi, quien manifiesta que en cumplimiento a la Ley N° 27470, por el cual establecen normas complementarias para la ejecución del programa del vaso de leche, solicitando la Conformación del Comité de Administración del Programa del Vaso de Leche, paralo cual, hace alcance de la propuesta según lo establecido en la norma con el siguiente detalle:\nITEM NOMBRE y APELLIDOS CARGO REPRESENTACION 01 MVZ. ADRIEL K. CARRILLO CAJIGAS Presidente Alcalde 02 Lie. ALBERTO FUENTES VEGA Miembro Sub Gerente Programas Sociales 03 Nutricionista MARILY TICONA JOSEC Miembro Centro de Salud Calca 04 Ing. WASSHINGTON PEÑA CARDENAS Miembro Agencia Agraria Calca 05 BRIZA1DA ZEVALLOS YUPANQUI Miembro Representante del Comité PVL 06 YESICA CUBA QUISPE Miembro Representante del Comité PVL 07 SAYDA MERMA PEREZ Miembro Representante del Comité PVL\n3- .9. Que, mediante Acta de Administración de Comité de fecha 18 de Febrero del año 2022, convocada por la Presidenta de los Comités del vaso de leche, acuerdan y aprueban a los nuevos miembros representantes de los comités del vaso del que deben de integrar el Comité de Administración del Programa del Vaso de Leche para el año 2022;\n3- .10. Que, teniendo en consideración que el Comité de Administración del Programa del Vaso de Leche, constituye un órgano directriz, encargado de cumplir y hacer cumplir el conjunto de acciones y disposiciones legales, que tienen como objetivo y finalidad del Programa del Vaso de Leche, se hace necesario formalizar su conformación con el dispositivo legal municipal, por lo que conforme al marco normativo vigente, es procedente la formalización de la Conformación, debiéndose previamente poner a consideración del Concejo

```
Municipal para su aprobación y/o ratificación según corresponda, para luego emitir la respectiva Resolución de Alcaldía;\nV. CONCLUSIÓN.\n3.1.- Por las consideraciones expuestas la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca CONCLUYE que es PROCEDENTE LA CONFORMACION DEL COMITÉ DE ADMINISTRACION DEL PROGRAMA DEL VASO\nDE LECHE - 2022, de la Municipalidad Provincial de Calca, la misma que estará integrada de acuerdo al siguiente detalla:\nITEM NOMBRE y APELLIDOS CARGO REPRESENTACION 01 MVZ. ADRIEL K. CARRILLO CAJIGAS Presidente Alcalde 02 Lic. ALBERTO FUENTES VEGA Miembro Sub Gerente Programas Sociales 03 Nutricionista MARILY TICONA JOSEC Miembro Centro de Salud Calca 04 Ing. WASSHINGTON PEÑA CARDENAS Miembro Agencia Agraria Calca 05 BRIZAIDA ZEVALLOS YUPANQUI Miembro Representante del Comité PVL 06 YESICA CUBA QUISPE Miembro Representante del Comité PVL 07 SAYDA MERMA PEREZ Miembro Representante del Comité PVL\nEl mismo que previo al acto resolutivo (Resolución de Alcaldía) debe ponerse a consideración del Concejo Municipal para su aprobación.\nV. RECOMENDACIONES.\n5.1.- Remitir el presente Expediente Administrativo a la Oficina de Secretaria General, a efecto de que se programa en Agenda para Sesión de Concejo y consideración de los miembros del Concejo Municipal.\n5.2.- Debe dejarse sin efecto de manera expresa, cualquier otro dispositivo de contravenga o se oponga a la implementación de la presente.\nSe adjunta al presente:\n- EXPEDIENTE con Reg. N° 1300-GM-MPC-2022, con Nueve (09) folios.\nAtentamente.",
  "input": "",
  "instruction": "RECONOCIMIENTODEL COMITÉ DE ADMINISTRACION DEL PROGRAMA VASO DE LECHE PARA EL AÑO 2022."
 },
]
```


4.4 Entrenamiento de modelos Llama y Mistral

El entrenamiento de los modelos LLaMA 3.1 y Mistral v0.3 se realizó mediante técnicas de fine-tuning supervisado, con el objetivo de adaptar los modelos al dominio jurídico específico.

4.4.1 Tokenización

La tokenización se realizó mediante el tokenizador oficial de LLaMA 3.1, utilizando la librería Unsloth. Se establecieron los siguientes parámetros:

- Longitud máxima de secuencia (max_seq_length): 8192 tokens.
- Batch size efectivo: 4 ejemplos por paso (considerando acumulación de gradientes).
- Token de fin de secuencia (EOS_TOKEN): agregado de manera explícita para evitar bucles infinitos en las generaciones.

La tokenización permitió transformar el dataset legal en un formato compatible con el modelo, asegurando que cada ejemplo se representara como una secuencia de vectores numéricos.

```
EOS_TOKEN = tokenizer.eos_token # Must add EOS_TOKEN
def formatting_prompts_func(examples):
    instructions = examples["instruction"]
    inputs       = examples["input"]
    outputs      = examples["output"]
    texts = []
    for instruction, input, output in zip(instructions, inputs, outputs):
        # Must add EOS_TOKEN, otherwise your generation will go on forever!
        text = alpaca_prompt.format(instruction, input, output) + EOS_TOKEN
        texts.append(text)
    return { "text" : texts, }
pass
#importar dataset
from datasets import Dataset
from datasets import load_dataset
dataset = Dataset.from_list(data)
dataset = dataset.map(formatting_prompts_func, batched = True,)
```

4.4.2 Adaptación del Modelo mediante LoRA

Para el ajuste fino del modelo se utilizó Low-Rank Adaptation (LoRA), técnica que permite entrenar modelos de gran tamaño de manera eficiente, actualizando solo un subconjunto reducido de parámetros.

Los valores empleados fueron:

- Rango de descomposición (r) = 16.
- Escala de adaptación (lora_alpha) = 16.
- Dropout aplicado a LoRA (lora_dropout) = 0.

Esta configuración permitió reducir significativamente el uso de memoria GPU y el tiempo de entrenamiento, manteniendo un rendimiento competitivo frente a un ajuste fino completo.

```
model = FastLanguageModel.get_peft_model(
    model,
    r = 16, # Choose any number > 0 ! Suggested 8, 16, 32, 64, 128 hace referencia al rango (rank) en el método LoRA (Low-Rank Adaptation)
    target_modules = ["q_proj", "k_proj", "v_proj", "o_proj",
                      "gate_proj", "up_proj", "down_proj",],
    lora_alpha = 16,#lora_alpha es un coeficiente de escalado que ajusta cuánto impacto tienen las matrices de bajo rango (r)
    lora_dropout = 0, # Supports any, but = 0 is optimized - evita sobreajuste
    bias = "none",    # Supports any, but = "none" is optimized - bias (sesgo) es un término adicional en las capas lineales (o convolucionales) que permite al modelo aprender desplazamientos en los datos
    # [NEW] "unsloth" uses 30% less VRAM, fits 2x larger batch sizes!
    use_gradient_checkpointing = "unsloth", # True or "unsloth" for very long context
    random_state = 3407,#se refiere a una semilla aleatoria (random seed) que se utiliza para asegurar que los procesos aleatorios
    use_rslora = False,  # We support rank stabilized LoRA
    loftq_config = None, # And LoftQ
)
```

4.4.3 Fine-Tuning supervisado (SFT) con trl.SFTTrainer

Para adaptar el modelo base (Llama 3.1 8B) al dominio jurídico se empleó un procedimiento de *Supervised Fine-Tuning (SFT)* mediante la librería trl (TRL: Transformer Reinforcement Learning utilities), concretamente la clase SFTTrainer. El uso de SFTTrainer permite estructurar el bucle de entrenamiento sobre un dataset textual formateado, con soporte directo para adaptadores LoRA y optimizadores de baja memoria.

4.4.4 Preparación del dataset para SFT

- El dataset se mantiene en formato JSON previamente explicado y se convierte a un objeto datasets.Dataset de Hugging Face.
- Cada registro contiene un campo textual único (en nuestro caso "text"), que integra el prompt + la respuesta (formato *alpaca*/instruction-following):
- { "text": "<PROMPT TEMPLATE + RESPUESTA esperada>“ }

- Se calculan estadísticas de tokens por ejemplo (mínimo, media, máximo) con el tokenizador del modelo para decidir la política de truncamiento o segmentación (split en bloques si excede max_seq_length).

4.4.5 Configuración de SFTTrainer

```
from trl import SFTConfig, SFTTrainer
trainer = SFTTrainer(
    model = model,
    tokenizer = tokenizer,
    train_dataset = dataset,
    dataset_text_field = "text",
    max_seq_length = max_seq_length,
    dataset_num_proc = 2,
    packing = False, # Can make training 5x faster for short sequences.
    args = SFTConfig(
        per_device_train_batch_size = 2,
        gradient_accumulation_steps = 4,
        warmup_steps = 5,
        # num_train_epochs = 1, # Set this for 1 full training run.
        max_steps = 40,
        learning_rate = 2e-4,
        logging_steps = 1,
        optim = "adamw_8bit",
        weight_decay = 0.01,
        lr_scheduler_type = "linear",
        seed = 3407,
        output_dir = "outputs",
        report_to = "none", # Use this for WandB etc
    ),
)
```

Parámetros principales y su justificación

- **model, tokenizer**: modelo (con adaptadores LoRA ya aplicados) y su tokenizador.
- **train_dataset**: objeto datasets.Dataset conteniendo los ejemplos (campo "text").
- **dataset_text_field = "text"**: nombre de la columna que contiene el texto a entrenar.
- **max_seq_length**: longitud máxima de tokens por ejemplo (ej. 8192). Debe concordar con la ventana de contexto del modelo y el preprocesamiento (división/packing).

- **dataset_num_proc = 2**: paralelismo para map/procesos de tokenización. Ajustar según CPU disponible.
- **packing = False**: si True, empaqueta varios ejemplos cortos en una misma secuencia para llenar la longitud máxima (incrementa eficiencia de cómputo y throughput para ejemplos cortos). Se dejó False en la configuración inicial para mayor trazabilidad; puede activarse para entrenamientos más largos/eficientes.
    - **SFTConfig (argumentos relevantes)**
- **per_device_train_batch_size = 2**: batch por GPU. Bajo por la limitación de VRAM al usar 4-bit quantization + LoRA.
- **gradient_accumulation_steps = 4**: acumula gradientes para simular un *effective batch size* de 2 * 4 = 8. Técnica habitual cuando mem. GPU es limitada.
- **warmup_steps = 5**: pasos de *warmup* del scheduler para estabilizar comienzo del entrenamiento.
- **max_steps = 40**: número total de pasos de entrenamiento (temporal/ejemplo). En un experimento real aumentarlo sustancialmente (p. ej. decenas de miles) o usar num_train_epochs. En tesis, documenta por qué escogiste un número.
- **learning_rate = 2e-4**: tasa de aprendizaje inicial para LoRA; valor típico para adaptadores LoRA (ajustable).
- **logging_steps = 1**: frecuencia de logs (puede reducirse para menos ruido).
- **optim = "adamw_8bit"**: optimizador compatible con bitsandbytes para ahorrar memoria (8-bit optimizer).
- **weight_decay = 0.01**: regularización.
- **lr_scheduler_type = "linear"**: scheduler lineal común en SFT.
- **seed = 3407**: semilla para reproducibilidad.

- **output_dir = "outputs"**: carpeta de checkpoints y registros.
- **report_to = "none"**: sin reporte a sistemas externos (WandB) — cambiar si quieres tracking.

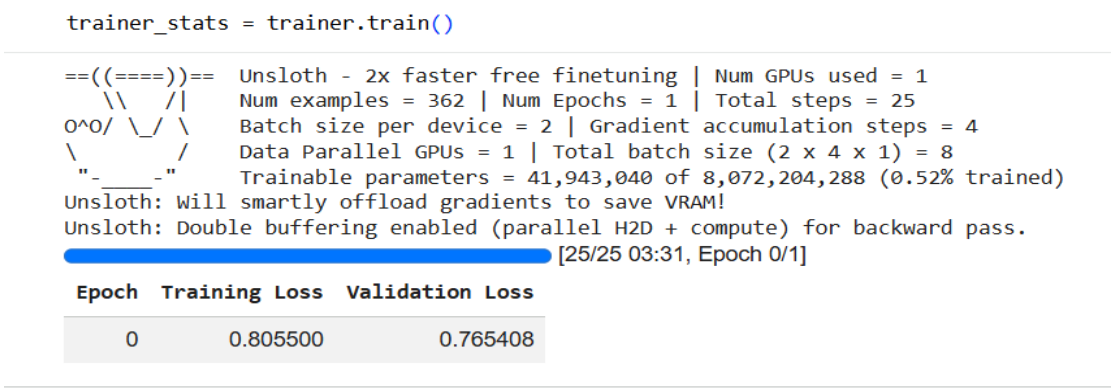
4.4.6 Flujo de entrenamiento

Una vez configurado el modelo y definidos los hiperparámetros de entrenamiento, se instanció el objeto **SFTTrainer** y se inició el proceso de ajuste fino mediante la siguiente instrucción:

```
trainer_stats = trainer.train()
```

Durante esta etapa, el **Trainer** se encargó de gestionar automáticamente el procesamiento por lotes (*batching*), la tokenización de los datos de entrenamiento (cuando correspondía), la ejecución de los procesos *forward* y *backward*, la actualización de los pesos entrenables y la administración de los puntos de control (*checkpoints*) del modelo.

Figura 4. *Ejecución del proceso de entrenamiento del modelo Llama 3.1*



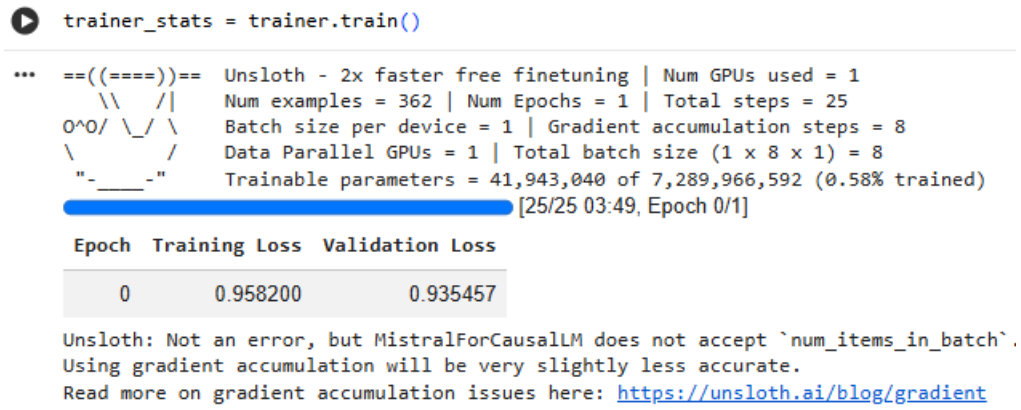
La Figura 7 muestra la ejecución del proceso de entrenamiento del modelo mediante la biblioteca Unsloth, observándose la configuración empleada y los principales indicadores obtenidos durante el ajuste fino. El entrenamiento se realizó utilizando 362 ejemplos, durante

una época (Epoch = 1), con un total de 25 pasos de entrenamiento, un tamaño de lote efectivo (*batch size*) de 8 y una GPU para el procesamiento.

Asimismo, se observa que únicamente fueron entrenados 41 943 040 parámetros, equivalentes al 0.52 % del total de 8 072 204 288 parámetros del modelo base, debido al empleo de la técnica LoRA (Low-Rank Adaptation), la cual permite realizar un ajuste fino eficiente reduciendo considerablemente el consumo de memoria y el tiempo de entrenamiento.

Finalmente, el proceso obtuvo una pérdida de entrenamiento (Training Loss) de 0.8055 y una pérdida de validación (Validation Loss) de 0.7654, evidenciando un comportamiento estable durante el ajuste del modelo y una adecuada capacidad de generalización sobre los datos de validación.

Figura 5. *Ejecución del proceso de entrenamiento del modelo*



4.4.7 Generación de Ejemplos

Con el modelo ajustado se procedió a la fase de inferencia, configurando los siguientes parámetros de decodificación:

- Temperature = 0.7: nivel de aleatoriedad controlada.
- Top-p = 0.9: muestreo nuclear para restringir el espacio de tokens probables.
- Máximo de tokens generados = 3300.

Se generaron ejemplos de opiniones legales, entre ellos un caso práctico sobre la implementación de un Plan de Gobierno Digital, en el cual el modelo fue capaz de reproducir la estructura obligatoria de antecedentes, base legal, análisis, conclusiones y recomendaciones.

4.4.8 Guardado y Reutilización del Modelo

El modelo resultante fue guardado en formato **LoRA adaptador**, utilizando los comandos:

- model.save_pretrained("lora_model")
- tokenizer.save_pretrained("lora_model")

Este enfoque permitió almacenar únicamente las matrices de adaptación entrenadas, lo que reduce el tamaño del modelo ajustado y facilita su reutilización en fases posteriores de evaluación.

4.5 Técnicas de ingeniería de prompts aplicado a la investigación

La ingeniería de prompts constituye una técnica fundamental en la utilización de modelos de lenguaje de gran escala (LLMs), ya que permite diseñar instrucciones estructuradas con el fin de guiar el comportamiento del modelo hacia la generación de resultados más precisos, coherentes y alineados con el contexto requerido. Esta disciplina ha adquirido relevancia debido a que pequeñas variaciones en la formulación del prompt pueden influir significativamente en la calidad de las respuestas generadas (Liu et al., 2023).

En el contexto de la presente investigación, las técnicas de ingeniería de prompts se emplean para optimizar la generación automática de opiniones legales, asegurando que los textos producidos por los modelos mantengan coherencia argumentativa, estructura formal adecuada y correspondencia con el lenguaje jurídico y las normativas vigentes. Esto resulta especialmente importante en el ámbito del derecho, donde la precisión terminológica y la correcta interpretación normativa son elementos críticos

De acuerdo con el repositorio Prompt Engineering Techniques Hub, las técnicas de ingeniería de prompts pueden clasificarse en diversas categorías, tales como técnicas básicas, técnicas de razonamiento, estrategias de descomposición del problema, mecanismos de autoevaluación y enfoques multietapa (KalyanKS-NLP, 2024). Estas categorías agrupan más de 25 técnicas diseñadas para mejorar el desempeño de los modelos de lenguaje en tareas complejas.

En este sentido, la presente investigación selecciona y aplica las técnicas Plan and Solve y Chain of Verification Prompting.

4.5.1 Técnica Plan-and-Solve Prompting

La técnica *Plan-and-Solve Prompting* es una estrategia de ingeniería de *prompts* que consiste en dividir la resolución de una tarea en dos fases claramente diferenciadas: en la primera, el modelo genera un plan estructurado para abordar el problema; y en la segunda, ejecuta dicho plan para producir la respuesta final. Este enfoque permite mejorar la organización del razonamiento y reducir errores en tareas complejas (Wang et al., 2023).

A diferencia de otras técnicas como *Chain of Thought*, que se centra en el razonamiento paso a paso, *Plan-and-Solve* introduce explícitamente una etapa de planificación previa, lo que resulta especialmente útil en problemas que requieren una estructura definida, como la redacción de documentos formales o informes técnicos.

En el presente trabajo de investigación, esta técnica fue aplicada en dos momentos consecutivos del proceso: primero, en la extracción y organización de antecedentes; y posteriormente, en la generación automatizada del informe legal. Esta secuencia responde a la lógica de funcionamiento del sistema, debido a que la correcta identificación de antecedentes constituye el insumo principal para la elaboración posterior del informe jurídico.

4.5.1.1 Aplicación en la extracción de antecedentes

Fase de planificación:

En esta fase, el modelo realiza internamente las siguientes acciones:

- Identifica los documentos relevantes ingresados al sistema.
- Reconoce fechas, remitentes, destinatarios, números de documento y asunto principal.
- Determina qué información resulta jurídicamente relevante para el caso.
- Clasifica los antecedentes según su importancia y orden cronológico.
- Elimina información repetitiva, irrelevante o no vinculada con la solicitud principal.

Fase de ejecución:

Luego de organizar la información, el sistema redacta los antecedentes de forma cronológica, objetiva y clara, respetando el orden de ocurrencia de los hechos y manteniendo un estilo institucional.

De esta manera, el modelo produce una sección de antecedentes estructurada, precisa y coherente, que posteriormente servirá como base para la generación del informe legal.

```
# ================= PROMPT ANTECEDENTES =================
PROMPT_ANTECEDENTES = """
Actúa como abogado especialista en asesoría jurídica administrativa.

OBJETIVO:
Extraer y organizar los ANTECEDENTES relevantes de un expediente administrativo
a partir de los actuados contenidos en el texto.

FASE 1 – IDENTIFICACIÓN:
Identifica informes, oficios, memorandos, resoluciones u otros documentos relevantes.

FASE 2 – EXTRACCIÓN:
Para cada documento identificado, extrae:
- Tipo de documento
- Número
- Oficina de origen
- Oficina de destino
- Asunto
- Fecha
- Descripción objetiva

FASE 3 – ORGANIZACIÓN:
Ordena los antecedentes cronológicamente y redáctalos en lenguaje técnico-jurídico,
formal e impersonal.

REGLAS:
- No resumas ni interpretes.
- No inventes información.
- Mantén fidelidad documental.
"""
```

Ilustración 1 Planificación para extraer antecedentes

4.5.1.2 Aplicación en la generación del informe legal

Fase de planificación

En la generación del informe legal, el modelo organiza internamente la información antes de redactar la respuesta final. Para ello:

- Determina la naturaleza jurídica del caso, identificando si se trata de contratación pública, obras públicas, ampliación de plazo, modificación presupuestal, procedimiento administrativo, gestión institucional, entre otros.

- Identifica palabras clave y conceptos jurídicos relevantes presentes en los antecedentes.
- Selecciona las normas aplicables de acuerdo con la materia identificada, priorizando la Constitución Política del Perú, leyes, reglamentos, directivas y disposiciones administrativas vigentes.
- Define la estructura lógica del informe legal, organizando la información en antecedentes, base legal, análisis, conclusiones, recomendaciones y jurisprudencia aplicable.
- Identifica los problemas jurídicos relevantes y establece la secuencia argumentativa que seguirá el análisis.

Esta etapa se desarrolla de manera interna y no aparece en la respuesta final generada al usuario. En el código implementado, esta lógica puede observarse en los prompts diseñados para cada sección del informe, donde se establece expresamente una "FASE 1 – PLANIFICACIÓN" previa a la redacción.

Fase de ejecución

Una vez definido el plan, el modelo procede a redactar el contenido del informe legal siguiendo estrictamente la estructura prevista. En esta fase:

- Redacta los antecedentes de manera cronológica y objetiva.
- Genera la Base Legal listando únicamente las normas pertinentes al caso.
- Desarrolla el análisis jurídico vinculando hechos y normas aplicables.
- Formula conclusiones coherentes con el análisis realizado.
- Emite recomendaciones dirigidas a las áreas competentes.
- Identifica jurisprudencia relevante del Tribunal Constitucional, Corte Suprema, Tribunal de Contrataciones del Estado y Contraloría General de la República.

A continuación, se verifica el prompt para generar la base legal:

```
# ================ PROMPTS ================
PROMPT_BASE_LEGAL = """
Actúa como abogado especialista en Derecho Administrativo peruano,
con amplia experiencia en la elaboración de opiniones legales institucionales
en el sector público.

OBJETIVO:
Elaborar exclusivamente la sección II. BASE LEGAL de un informe legal,
limitándote a IDENTIFICAR y LISTAR las normas jurídicas aplicables,
utilizando únicamente los ANTECEDENTES proporcionados en formato JSON
y conforme al ordenamiento jurídico vigente en el Perú.

FASE 1 – PLANIFICACIÓN (Plan and Solve):
A partir de los antecedentes, identifica internamente la naturaleza
jurídico-administrativa del caso (por ejemplo: obra pública, inversión
pública, contratación estatal, gestión administrativa, modificación
presupuestal, ampliación de plazo u otros).
Esta identificación es solo para orientar la selección normativa y
NO debe ser incluida en la respuesta.

FASE 2 – SELECCIÓN NORMATIVA:
Determina únicamente las normas pertinentes al caso identificado,
priorizando:
- Constitución Política del Perú (solo si resulta aplicable)
- Leyes y Decretos Legislativos
- Reglamentos
- Directivas, resoluciones o disposiciones administrativas vigentes

FASE 3 – REDACCIÓN DE LA BASE LEGAL:
Redacta la sección II. BASE LEGAL LIMITÁNDOTE EXCLUSIVAMENTE a:
- Listar las normas aplicables
- Indicar el nombre completo de la norma
- Señalar su número y año de promulgación (cuando corresponda)

NO debes:
- Explicar el contenido de las normas
- Describir artículos
- Relacionar normas con los antecedentes
- Realizar análisis, interpretaciones ni conclusiones

CHAIN OF VERIFICATION (Verificación obligatoria antes de responder):
- Verifica que todas las normas citadas existan en el ordenamiento jurídico peruano.
- Verifica que las normas estén vigentes o sean aplicables al caso.
- Elimina cualquier norma irrelevante o no sustentada en los antecedentes.
- Asegura que no se haya incluido ninguna descripción normativa.
```

```
REGLAS ESTRICTAS:
- No inventes normas, artículos ni disposiciones.
- No incluyas artículos ni incisos.
- No utilices frases explicativas.
- Usa lenguaje técnico-jurídico, formal e impersonal.
- Redacta en formato institucional.

FORMATO DE SALIDA OBLIGATORIO:

II. BASE LEGAL

- [Listado enumerado o con viñetas de normas jurídicas]
"""
```

4.5.2 Técnica Chain of Verification Prompting

La técnica *Chain of Verification* es un enfoque avanzado de ingeniería de *prompts* que consiste en generar una respuesta inicial y posteriormente someterla a un proceso de verificación mediante una serie de preguntas o criterios evaluativos. Esta técnica busca mejorar la fiabilidad y precisión de las respuestas generadas por modelos de lenguaje, especialmente en contextos donde la exactitud es crítica (Dhuliawala et al., 2023).

Su principal finalidad es reducir el fenómeno conocido como hallucination, entendido como la generación de información aparentemente válida pero incorrecta, inexistente o no sustentada. En contextos jurídicos, este problema resulta especialmente crítico debido a que una referencia normativa errónea, una conclusión inconsistente o una interpretación imprecisa puede afectar directamente la validez del documento generado.

A diferencia de otras técnicas centradas únicamente en la producción de respuestas, Chain of Verification incorpora un mecanismo de autocontrol que obliga al modelo a revisar el contenido antes de emitir la respuesta final. De esta manera, se mejora la calidad de los

resultados y se disminuye la probabilidad de incluir normas inexistentes, hechos no acreditados o argumentos contradictorios.

En la presente investigación, esta técnica fue utilizada como una etapa complementaria al Plan-and-Solve Prompting, tanto en la extracción de antecedentes como en la generación del informe legal. Su aplicación permitió verificar la consistencia entre antecedentes, base legal, análisis, conclusiones y recomendaciones, asegurando que cada sección del informe mantenga coherencia lógica y jurídica.

La técnica Chain of Verification fue implementada mediante reglas de validación incorporadas en los prompts del sistema. Estas verificaciones se aplicaron antes de generar la respuesta final, permitiendo que el modelo revise internamente la calidad del contenido producido.

4.5.2.1 Revisión de coherencia lógica

En esta etapa, el modelo verifica que exista relación lógica entre los antecedentes, el análisis jurídico, las conclusiones y las recomendaciones.

Esto implica revisar que:

- Las conclusiones deriven directamente del análisis realizado.
- Las recomendaciones respondan a las conclusiones previamente formuladas.
- No existan contradicciones entre distintas partes del informe.
- La secuencia argumentativa mantenga orden y coherencia.

Por ejemplo, si en el análisis se determina que una actuación administrativa incumple una norma específica, la conclusión debe reflejar dicha situación y la recomendación debe orientarse a corregirla o prevenir responsabilidades.

4.5.2.2 Validación normativa

La validación normativa constituye uno de los aspectos más relevantes de esta técnica, debido a que el informe legal debe sustentarse únicamente en normas vigentes y pertinentes al caso.

En esta fase, el modelo verifica que:

- Todas las normas citadas existan en el ordenamiento jurídico peruano.
- Las normas seleccionadas sean aplicables a la materia identificada.
- No se incluyan normas derogadas, impertinentes o inexistentes.
- No se incorporen artículos, reglamentos o disposiciones no mencionadas previamente en la Base Legal.
- Exista correspondencia entre la norma utilizada y el argumento jurídico desarrollado.

En el código implementado, esta validación se observa principalmente en la sección de Base Legal, donde se instruye al modelo a "verificar que todas las normas citadas existan en el ordenamiento jurídico peruano" y "eliminar cualquier norma irrelevante o no sustentada en los antecedentes".

4.5.2.3 Consistencia del razonamiento jurídico

La técnica también permite verificar que el razonamiento jurídico mantenga consistencia interna y se sustente exclusivamente en los hechos acreditados en los antecedentes.

Para ello, el modelo revisa que:

- No se incorporen hechos nuevos no contenidos en los antecedentes.
- Cada argumento jurídico tenga respaldo en una norma previamente identificada.
- No se introduzcan interpretaciones contradictorias.
- El análisis no anticipe conclusiones ni recomendaciones.
- Las conclusiones no repitan textualmente el análisis ni introduzcan elementos nuevos.

Esta validación es especialmente importante porque evita que el modelo genere afirmaciones no sustentadas o desarrolle argumentos alejados del caso concreto.

4.5.2.4 Adecuación del lenguaje jurídico

Finalmente, la técnica Chain of Verification también se empleó para controlar la forma en que se redacta el informe legal.

En esta fase, el modelo verifica que:

- Se utilice lenguaje técnico-jurídico formal e institucional.
- Los párrafos del análisis inicien con la expresión "Que,", según la estructura tradicional de las opiniones legales administrativos.
- No se utilicen expresiones coloquiales, ambiguas o informales.
- Las recomendaciones comiencen obligatoriamente con la expresión "Se recomienda".
- El documento mantenga uniformidad terminológica y estilística.

Este control resulta importante porque, además de asegurar la precisión jurídica, permite que el informe generado conserve el estilo propio de los documentos emitidos por oficinas de asesoría jurídica del sector público.

4.6 Ejecución Experimental

La ejecución experimental tuvo como finalidad evaluar el desempeño de los modelos de lenguaje LLaMA 3.1 (8B), Mistral v0.3 (7B) y Claude Haiku en la generación automatizada de opiniones legales para la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca. La evaluación consideró tanto la calidad técnica de los documentos generados como la capacidad de cada modelo para producir textos jurídicamente coherentes, estructurados y consistentes con la normativa vigente.

Con este propósito se definieron dos escenarios experimentales:

Modelos ajustados mediante Fine-Tuning: LLaMA 3.1 (8B) y Mistral v0.3 (7B), utilizando la técnica LoRA (Low-Rank Adaptation) y el framework SFTTrainer.

Modelo basado en ingeniería de prompts: Claude Haiku, evaluado mediante inferencia directa a través de API empleando las estrategias Plan-and-Solve Prompting y Chain of Verification Prompting, sin requerir entrenamiento adicional.

Para los modelos LLaMA y Mistral se utilizó el dataset jurídico construido durante la investigación, compuesto por opiniones legales emitidas por la Municipalidad Provincial de Calca. Ambos modelos fueron entrenados utilizando el mismo formato de entrada y salida, garantizando condiciones homogéneas durante el ajuste fino. Posteriormente, se realizó la fase de inferencia utilizando expedientes diferentes a los empleados durante el entrenamiento, con el objetivo de evaluar la capacidad de generalización del modelo.

Por su parte, Claude Haiku fue evaluado mediante inferencia directa, utilizando instrucciones especializadas diseñadas para reproducir la estructura utilizada por la Oficina de Asesoría Jurídica.

Con el propósito de garantizar una comparación homogénea, los tres modelos fueron sometidos a los mismos casos de prueba, correspondientes a expedientes reales relacionados con reconocimiento de organizaciones sociales, contrataciones públicas, modificaciones presupuestales, ampliaciones de plazo, obras públicas, gestión institucional y procedimientos administrativos.

Cada modelo recibió exactamente la misma información de entrada y se evaluó su capacidad para generar las secciones obligatorias de una opinión legal:

- Antecedentes.
- Base legal.

- Análisis jurídico.
- Conclusiones.
- Recomendaciones.

Durante la fase de inferencia se utilizaron los siguientes parámetros de generación:

- Temperature = 0.7.
- Top-p = 0.9.
- Máximo de tokens generados = 3300.
- Penalización de frecuencia reducida.
- Longitud máxima del contexto adaptada a cada modelo.

Asimismo, durante la ejecución experimental se registraron indicadores relacionados con el desempeño del sistema, entre ellos:

- Calidad semántica de las respuestas.
- Tiempo promedio de inferencia.
- Cantidad de tokens generados.
- Consumo aproximado de memoria.
- Coherencia estructural.
- Precisión normativa.
- Completitud de las secciones requeridas.

4.6.1 Modelo Llama 3.1

La ejecución experimental del modelo LLaMA 3.1 (8B) tuvo como finalidad evaluar su capacidad para generar opiniones legales de manera automatizada a partir del dataset jurídico construido para la presente investigación. Para ello, el modelo fue sometido a un proceso de ajuste fino supervisado (fine-tuning) mediante la técnica LoRA (Low-Rank Adaptation) y la

herramienta SFTTrainer, utilizando los documentos jurídicos previamente estructurados bajo el formato instrucción–entrada–salida.

Una vez finalizado el proceso de entrenamiento, se procedió a la fase de inferencia utilizando casos de prueba no empleados durante el entrenamiento, con la finalidad de evaluar la capacidad del modelo para generar opiniones legales a partir de nuevos expedientes. La generación se realizó empleando los parámetros establecidos durante la configuración experimental, entre ellos una temperatura de 0.7, Top-p de 0.9 y un máximo de 3300 tokens generados.

Las opiniones generadas fueron estructuradas considerando las principales secciones requeridas para un documento jurídico, tales como antecedentes, base legal, análisis jurídico, conclusiones y recomendaciones. De esta manera, se buscó verificar no solo la capacidad del modelo para producir texto, sino también su capacidad para mantener una estructura formal y desarrollar contenido jurídicamente coherente.

Para evaluar el desempeño de LLaMA 3.1 se consideraron tres grupos principales de indicadores: métricas de calidad del contenido generado, tiempo promedio de inferencia y cantidad promedio de tokens generados. Estos indicadores permitieron analizar tanto la calidad de las opiniones legales como el comportamiento computacional del modelo durante su ejecución.

A partir de esta ejecución experimental se desarrollaron los siguientes aspectos:

4.6.1.1 Métricas de calidad del modelo Llama 3.1.

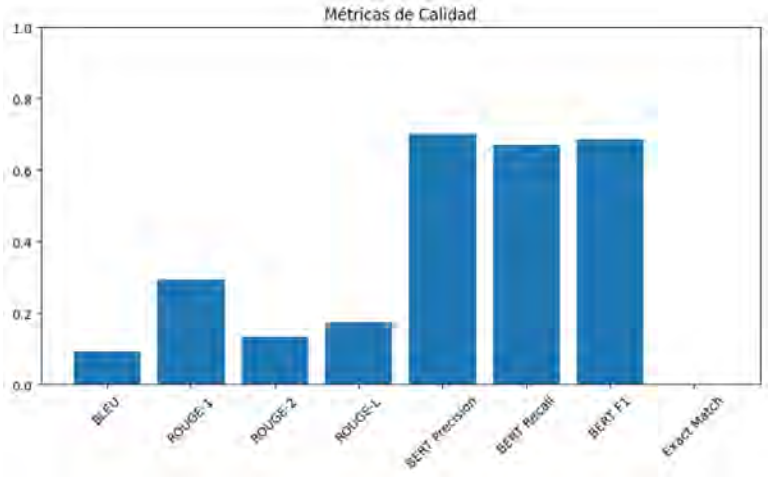
Con el propósito de evaluar objetivamente la calidad de las opiniones legales generadas, se emplearon métricas ampliamente utilizadas en tareas de procesamiento del lenguaje natural, tales como BLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERT Precision, BERT Recall, BERT F1 y Exact Match.

Los resultados muestran que las métricas basadas en BERTScore alcanzaron los valores más altos (Precision = 0.70, Recall = 0.67 y F1 = 0.69), evidenciando una elevada similitud semántica entre las opiniones legales generadas por el modelo y los documentos elaborados por especialistas jurídicos.

Por otra parte, las métricas ROUGE obtuvieron valores de 0.29, 0.13 y 0.17, mientras que BLEU alcanzó un valor de 0.09, indicando que el modelo no reproduce literalmente los documentos originales, sino que conserva adecuadamente su contenido jurídico y argumentativo.

Finalmente, el valor de Exact Match igual a 0 resulta consistente con la naturaleza de la redacción jurídica, ya que existen múltiples formas correctas de expresar un mismo criterio legal sin que el texto sea idéntico.

Figura 6. *Métricas de calidad del modelo Llama 3.1.*



4.6.1.2 Tiempo promedio de inferencia Modelo Llama 3.1

Además de evaluar la calidad del texto generado, se analizó el rendimiento computacional del modelo durante la fase de inferencia. La Figura 9 muestra que el tiempo promedio requerido para generar una opinión legal fue de aproximadamente 106 segundos, incluyendo el procesamiento del expediente, la recuperación del contexto jurídico y la generación del documento completo.

Este resultado demuestra que el prototipo puede elaborar opiniones legales en un tiempo considerablemente menor al proceso tradicional desarrollado por un especialista jurídico, reduciendo significativamente el tiempo de respuesta y mejorando la eficiencia operativa de la Oficina de Asesoría Jurídica.

4.6.1.3 Cantidad promedio de tokens generados Modelo Llama 3.1

Durante la ejecución experimental también se registró la cantidad de texto producido por el modelo en cada inferencia. Como se observa en la Figura 10, el sistema generó aproximadamente 1 670 tokens por documento, longitud suficiente para desarrollar una opinión legal completa, incorporando antecedentes, base legal, análisis jurídico, conclusiones y recomendaciones.

La cantidad de tokens obtenida evidencia que el modelo mantiene un equilibrio entre la amplitud del razonamiento jurídico y la eficiencia computacional, permitiendo generar documentos técnicamente completos sin exceder los límites establecidos durante la configuración experimental.

En el caso específico de Claude Haiku, la ejecución experimental se desarrolló en dos etapas consecutivas: inicialmente se realizó la extracción automatizada de antecedentes contenidos en el expediente administrativo y posteriormente se generó la opinión legal utilizando dicha información como contexto de entrada. Esta metodología permitió reproducir

el flujo de trabajo seguido por los especialistas de la Oficina de Asesoría Jurídica, donde primero se analiza la documentación del expediente y posteriormente se desarrolla el razonamiento jurídico correspondiente.

Finalmente, todas las opiniones legales generadas fueron almacenadas y comparadas para identificar fortalezas, limitaciones y diferencias de comportamiento entre los modelos evaluados, proporcionando evidencia técnica y funcional para seleccionar el modelo con mejor desempeño en la generación automatizada de opiniones legales.

4.6.2 Modelo Mistral

La ejecución experimental del modelo Mistral v0.3 (7B) tuvo como finalidad evaluar su capacidad para generar opiniones legales a partir del dataset jurídico construido durante la investigación. Al igual que Llama 3.1, el modelo fue sometido a un proceso de fine-tuning supervisado mediante LoRA y SFTTrainer, utilizando el conjunto de datos estructurado bajo el formato instrucción–entrada–salida.

El proceso de evaluación se desarrolló bajo condiciones experimentales homogéneas, empleando los mismos casos de prueba y parámetros de generación utilizados para los demás modelos, con la finalidad de garantizar la comparabilidad de los resultados. La evaluación comprendió métricas de calidad textual y semántica, así como indicadores relacionados con el rendimiento computacional durante la fase de inferencia.

4.6.2.1 Métricas de calidad del modelo Mistral v0.3

Con el propósito de evaluar objetivamente la calidad de las opiniones legales generadas por el modelo Mistral v0.3, se emplearon métricas de evaluación automática utilizadas en tareas de procesamiento del lenguaje natural, entre ellas BLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERT Precision, BERT Recall, BERT F1 y Exact Match.

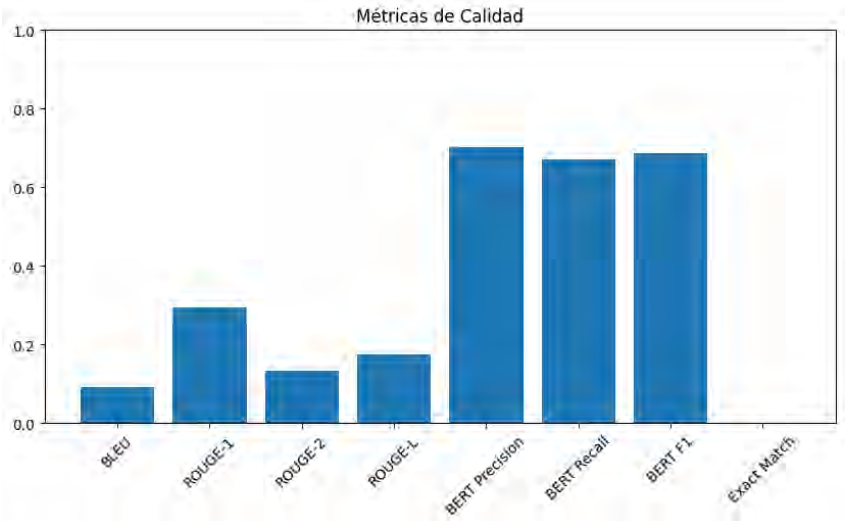
Estas métricas permitieron analizar diferentes dimensiones del contenido generado. Mientras que BLEU y ROUGE permiten determinar el grado de coincidencia léxica entre el texto generado y el documento de referencia, las métricas basadas en BERTScore permiten evaluar la similitud semántica, considerando el significado del contenido independientemente de que las expresiones utilizadas sean exactamente iguales.

Los resultados obtenidos evidencian el comportamiento del modelo Mistral v0.3 durante la generación de opiniones legales. En particular, las métricas basadas en BERTScore permiten determinar el nivel de correspondencia semántica alcanzado entre las opiniones generadas automáticamente y los documentos elaborados por especialistas jurídicos.

Por otra parte, los valores obtenidos mediante ROUGE permiten observar el nivel de coincidencia de términos y secuencias de palabras con respecto a los documentos de referencia, mientras que el valor de BLEU permite determinar el grado de similitud de las expresiones generadas. En este contexto, valores moderados en estas métricas no necesariamente implican una baja calidad jurídica, debido a que una opinión legal puede expresar correctamente un mismo razonamiento utilizando una redacción diferente a la del documento original.

Asimismo, el indicador Exact Match permite determinar si la respuesta generada coincide exactamente con el documento de referencia. En el contexto de la presente investigación, un valor bajo o igual a cero resulta esperable, debido a que el objetivo del modelo no consiste en reproducir literalmente una opinión legal existente, sino en generar una nueva opinión manteniendo su estructura, contenido jurídico y coherencia argumentativa.

Figura 7. *Métricas de calidad del modelo Mistral v0.3.*



En conjunto, las métricas obtenidas permiten establecer el nivel de similitud textual y semántica alcanzado por Mistral v0.3.

4.6.2.2 Tiempo promedio de inferencia del modelo Mistral v0.3

Además de evaluar la calidad del contenido generado, se analizó el rendimiento computacional del modelo Mistral v0.3 durante la fase de inferencia. Para ello, se registró el tiempo requerido por el modelo para procesar la información proporcionada y generar la opinión legal correspondiente.

El tiempo promedio de inferencia constituye un indicador importante dentro de la presente investigación, debido a que uno de los propósitos del uso de inteligencia artificial es contribuir a la optimización del proceso de elaboración de opiniones legales. La medición permitió determinar cuánto tiempo requiere el modelo para transformar la información contenida en el expediente en un documento jurídico estructurado.

Los resultados obtenidos permiten observar el comportamiento temporal del modelo durante la generación de las opiniones legales.

Asimismo, el tiempo registrado debe interpretarse considerando que la generación automática comprende diferentes procesos, entre ellos la preparación de la entrada, el procesamiento del contexto jurídico, la inferencia del modelo y la generación de los tokens correspondientes al documento final.

4.6.2.3 Cantidad promedio de tokens generados por el modelo Mistral v0.3

Durante la ejecución experimental también se registró la cantidad de tokens generados por Mistral v0.3 en cada proceso de inferencia. Este indicador permite determinar la extensión promedio de las opiniones legales generadas y analizar si el modelo es capaz de producir documentos suficientemente desarrollados para cubrir las diferentes secciones requeridas.

La cantidad de tokens generados constituye un indicador complementario del rendimiento del modelo, debido a que una mayor cantidad de tokens no necesariamente implica una mejor respuesta. Por ello, este indicador debe analizarse conjuntamente con las métricas de calidad y con la evaluación realizada por los especialistas jurídicos.

En el contexto de la presente investigación, se consideró que una cantidad adecuada de tokens debía permitir desarrollar las principales secciones de una opinión legal, comprendiendo antecedentes, base legal, análisis jurídico, conclusiones y recomendaciones, sin generar contenido innecesario o excesivamente repetitivo.

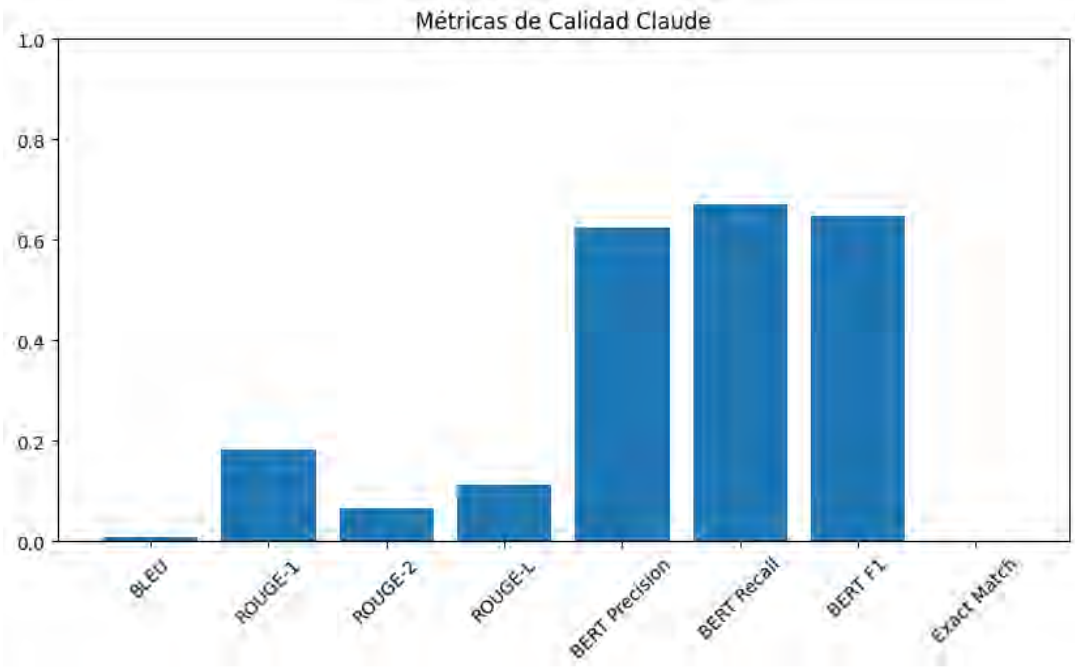
4.6.3 Modelo Claude

4.6.3.1 Métricas de calidad del modelo Claude

El modelo Claude presentó un BLEU promedio de 0,0099 y valores ROUGE entre 0,0668 y 0,1817, mientras que BERTScore F1 alcanzó al promedio 0,6469, evidenciando una similitud semántica moderada con los documentos de referencia. El Exact Match fue 0,0000, debido a la ausencia de coincidencia textual exacta. Asimismo, el tiempo promedio de inferencia fue de 33,81 segundos, con 4000 tokens generados aproximadamente. Estos

resultados permiten valorar tanto la similitud de las opiniones legales generadas como la eficiencia de respuesta del modelo.

Figura 8. *Tiempo promedio de inferencia del modelo Claude*



4.6.3.2 Tiempo promedio de inferencia del modelo Claude

Claude generó un promedio de 4.000 tokens por inferencia, lo que evidencia una producción extensa de contenido jurídico en cada generación.

4.6.3.3 Cantidad promedio de tokens generados por el modelo Claude

La inferencia se genera en un tiempo promedio de 33,81 segundos, evidenciando una capacidad de generación de un volumen considerable de contenido jurídico en un tiempo relativamente reducido.

4.7 Evaluación de rendimiento

La evaluación de rendimiento tuvo como finalidad comparar la capacidad de los modelos para generar opiniones legales con correspondencia respecto a los documentos de

referencia, considerando tanto la similitud del contenido generado como el tiempo requerido para su generación.

Se evaluaron tres configuraciones de modelos de lenguaje: Llama 3.1 y Mistral v0.3, adaptados mediante fine-tuning sobre el dataset jurídico, y Claude, utilizado mediante técnicas de ingeniería de prompts. La comparación se realizó utilizando métricas de similitud textual, similitud semántica y eficiencia temporal.

En cuanto a la similitud textual, se utilizaron BLEU, ROUGE-1, ROUGE-2 y ROUGE-L. BLEU permite medir la coincidencia de unidades léxicas (n-gramas) entre el texto generado y el texto de referencia, mientras que las métricas ROUGE permiten evaluar la coincidencia de diferentes unidades lingüísticas, considerando variantes como ROUGE-1, basada en unigramas; ROUGE-2, basada en bigramas; y ROUGE-L, basada en la subsecuencia común más larga(Papineni, 2002).

Por otro lado, se emplearon BERT Precision, BERT Recall y BERT F1 para evaluar la similitud semántica mediante representaciones contextuales. Estas métricas permiten identificar correspondencias de significado entre el texto generado y el texto de referencia, incluso cuando no existe una coincidencia exacta de palabras (Zhang et al., 2020). La utilización conjunta de métricas de similitud textual y semántica resulta pertinente, debido a que la coincidencia léxica por sí sola no necesariamente implica equivalencia de significado. Asimismo, el empleo de diferentes métricas permite obtener una evaluación más completa de la calidad de los textos generados, considerando distintas dimensiones del desempeño (Alexander R. Fabbri, 2021).

Como indicadores complementarios, se incorporaron el tiempo promedio de generación, el número promedio de tokens y Exact Match. El tiempo promedio de generación permitió valorar la eficiencia temporal de cada modelo, mientras que el número promedio de tokens permitió caracterizar la extensión de las respuestas generadas. Por su parte, Exact Match

permitió verificar si la respuesta generada coincidía de manera exacta con el texto de referencia. Sin embargo, debido a que los tres modelos obtuvieron un valor de cero en esta métrica, Exact Match no permitió establecer diferencias de desempeño entre ellos.

Tabla 3 *Comparación de modelos*

	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BERT Precision	BERT Recall	BERT Fl	Exact Match	Tiempo Promedio (s)	Tokens Promedio
CLAUDE	0.058738	0.28701	0.201403	0.169072	0.739016	0.735489	0.737248	0	33.81407	4000
MISTRAL	0.036796	0.159672	0.065791	0.106364	0.294822	0.588669	0.606974	0	114.0863	1831.066
LLAMA	0.091652	0.294274	0.134796	0.173982	0.703614	0.670019	0.68578	0	105.961	1673.582

Nota: Elaboración propia

Los resultados muestran que Llama 3.1 obtuvo los mayores valores en BLEU (0.0917), ROUGE-1 (0.2943) y ROUGE-L (0.1740). Por su parte, Claude presentó el mayor valor en ROUGE-2 (0.2014) y obtuvo los mejores resultados en BERT Precision (0.7390), BERT Recall (0.7355) y BERT F1 (0.7372). Mistral v0.3 presentó valores inferiores en las métricas de similitud consideradas. En cuanto al tiempo promedio de generación, Claude presentó el menor tiempo, con 33.81 segundos, frente a 105.96 segundos para Llama 3.1 y 114.09 segundos para Mistral v0.3.

Los resultados evidencian que ningún modelo obtuvo el mejor desempeño en todas las métricas. Llama 3.1 presentó una mayor coincidencia textual en algunas métricas, mientras que Claude presentó una mayor similitud semántica y una ventaja considerable en tiempo de generación. Por esta razón, la selección del modelo no se realizó mediante el criterio de mayor cantidad de métricas ganadas, sino mediante un procedimiento de evaluación multicriterio que permitiera integrar los diferentes aspectos del desempeño.

4.7.1 Criterio multicriterio para la selección del modelo

Para determinar el modelo con mejor desempeño global, los indicadores fueron agrupados en tres dimensiones: similitud léxica, similitud semántica y eficiencia temporal.

La dimensión de similitud textual estuvo conformada por BLEU, ROUGE-1, ROUGE-2 y ROUGE-L. La dimensión de similitud semántica estuvo conformada por BERT Precision, BERT Recall y BERT F1. Finalmente, la dimensión de eficiencia temporal estuvo representada por el tiempo promedio de generación.

Esta agrupación permitió evitar que una métrica individual determinara por sí sola la selección del modelo y permitió integrar diferentes características del proceso de generación de opiniones legales.

4.7.1.1 Asignación de pesos

Para determinar la importancia relativa de cada dimensión, se establecieron las siguientes ponderaciones:

Tabla 4 *Asignación de pesos*

Dimensión	Indicadores	Peso
Similitud semántica	BERT Precision, BERT Recall, BERT F1	**50%**
Similitud textual	BLEU, ROUGE-1, ROUGE-2, ROUGE-L	**35%**
Eficiencia temporal	Tiempo promedio de generación	**15%**
Total	—	**100%**

Nota: Elaboración propia

La asignación de los pesos se realizó con base en la relevancia de cada dimensión respecto del objetivo de la investigación. En particular, se otorgó la mayor ponderación a la similitud semántica debido a que, en la generación de opiniones legales, resulta prioritario conservar el significado y el contexto del documento de referencia, incluso cuando el contenido generado utilice palabras o estructuras sintácticas diferentes.

La similitud textual recibió una ponderación del 35 %, debido a que la coincidencia léxica constituye un indicador relevante de correspondencia con el documento de referencia, pero no necesariamente garantiza la conservación del significado. Finalmente, la eficiencia temporal recibió un peso del 15 %, debido a que representa un criterio importante para la viabilidad operativa del modelo, aunque la rapidez de generación no constituye por sí misma un indicador de calidad del contenido jurídico.

Por tanto, las ponderaciones utilizadas corresponden a un criterio metodológico definido para esta investigación y no a un estándar universal. Su finalidad fue reflejar la prioridad otorgada a la preservación del significado jurídico, sin excluir la correspondencia textual ni la eficiencia temporal.

4.7.2 Normalización de los indicadores

Debido a que los indicadores evaluados presentan diferentes escalas y unidades de medida, sus valores originales no fueron sumados directamente. Para hacerlos comparables, se aplicó una normalización min–max, transformando los resultados a una escala común entre 0 y 1, para las métricas en las que un valor mayor representa un mejor desempeño, se utilizó:

$$N_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}$$

Donde:

- N_{ij} representa el valor normalizado del modelo i en el indicador j;
- x_{ij} representa el valor original obtenido por el modelo i;
- $\min(x_j)$ representa el menor valor observado entre los modelos para el indicador j;
- $\max(x_j)$ representa el mayor valor observado entre los modelos para el indicador j.

Para el tiempo promedio de generación, debido a que un menor tiempo representa un mejor desempeño, se aplicó la normalización inversa:

$$N_{ij} = \frac{\max(x_{ij}) - x_j}{\max(x_j) - \min(x_j)}$$

De esta manera, todos los indicadores quedaron expresados en una escala común entre 0 y 1, donde un valor cercano a 1 representa un mejor desempeño relativo entre los modelos evaluados.

Es importante señalar que un valor normalizado de 0 no significa que el modelo no presente desempeño en el indicador original. Únicamente significa que obtuvo el menor valor relativo entre los modelos considerados para dicho indicador.

4.7.3 Cálculo de las dimensiones

Una vez normalizados los indicadores, se calculó el valor correspondiente a cada dimensión.

Para la similitud textual, se calculó el promedio de los cuatro indicadores:

$$ST_i = \frac{BLEU_i + ROUGE_i + ROUGE2_i + ROUGEL_i}{4}$$

Para la similitud semántica, se calculó el promedio de los tres indicadores:

$$SS_i = \frac{BP_i + BR_i + BF1_i}{3}$$

donde BP, BR y BF1 representan BERT Precision, BERT Recall y BERT F1, respectivamente.

Para la eficiencia temporal, se utilizó directamente el valor normalizado del tiempo promedio de generación:

$$ET_i = N_{Tiempo,i}$$

De esta manera, cada dimensión quedó representada por un único valor normalizado entre 0 y 1.

4.7.4 Cálculo del puntaje global

Una vez obtenidos los valores de cada dimensión, se calculó el puntaje global de cada modelo mediante la suma ponderada de las tres dimensiones:

$$PG_i = (0.35xST_i) + (0.50xSS_i) + (0.15xET_i)$$

donde:

PG_i = puntaje global del modelo i;

ST_i= similitud textual normalizada;

SS_i= similitud semántica normalizada;

ET_i = eficiencia temporal normalizada.

El modelo con el mayor puntaje global fue considerado como el de mejor desempeño bajo los criterios definidos en la investigación.

4.7.5 Resultados del análisis multicriterio

Aplicando las fórmulas con los datos de la tabla 3, se obtiene los siguientes resultados.

Tabla 5 *Resultados del análisis multicriterio*

Modelo	Similitud léxica	Similitud semántica	Eficiencia temporal	Puntaje global	Posición
Claude	0.8184	1	1	0.9364	1.º
Llama 3.1	0.8772	0.6932	0.1012	0.6688	2.º
Mistral v0.3	0	0	0	0	3.º

Nota: Elaboración propia

Los resultados del análisis multicriterio muestran que Claude obtuvo el mayor puntaje global, con un valor de 0.9364, seguido de Llama 3.1, con 0.6688, y Mistral v0.3, con 0.0000. Este resultado evidencia que la selección del modelo no depende exclusivamente de la cantidad de métricas individuales en las que cada modelo obtiene el primer lugar, sino del desempeño conjunto en las dimensiones establecidas y de la ponderación asignada a cada una.

Llama 3.1 obtuvo el mayor valor normalizado en la dimensión de similitud textual, con 0.8772, superando a Claude, que alcanzó 0.8184. Esto se encuentra relacionado con los mejores resultados obtenidos por Llama 3.1 en BLEU, ROUGE-1 y ROUGE-L. Sin embargo, Claude obtuvo el mayor desempeño en similitud semántica, con un valor normalizado de 1.0000, frente a 0.6932 obtenido por Llama 3.1. Asimismo, Claude alcanzó el máximo valor normalizado en eficiencia temporal, debido a que presentó el menor tiempo promedio de generación entre los modelos evaluados.

La ventaja de Claude en las dimensiones de similitud semántica y eficiencia temporal compensó su menor desempeño relativo en algunas métricas de similitud textual. Esto resulta consistente con las ponderaciones establecidas, dado que la similitud semántica recibió el mayor peso (50 %), seguida de la similitud textual (35 %) y la eficiencia temporal (15 %). En consecuencia, Claude alcanzó el mayor puntaje global y fue seleccionado como el modelo con mejor desempeño bajo los criterios definidos para esta investigación.

En el caso de Mistral v0.3, el valor normalizado de 0.0000 en las tres dimensiones no significa que sus resultados originales hayan sido iguales a cero. Este valor corresponde a su posición relativa como el modelo con los menores valores en los indicadores considerados para la comparación. Por tanto, el resultado debe interpretarse dentro del procedimiento de normalización aplicado y no como ausencia de capacidad de generación.

En consecuencia, el análisis multicriterio permitió integrar de manera sistemática la similitud textual, la similitud semántica y la eficiencia temporal, evitando seleccionar el modelo únicamente por el número de métricas en las que obtuvo el primer lugar. Bajo este procedimiento, Claude obtuvo el mayor puntaje global (0.9364), por lo que fue seleccionado como el modelo de mejor desempeño para la generación de opiniones legales en el contexto de la presente investigación.

4.7.6 Definición de las dimensiones de evaluación

A partir de los criterios utilizados en la evaluación multicriterio, se establecieron dos dimensiones principales para evaluar el rendimiento de los modelos: calidad y eficiencia.

La dimensión calidad corresponde al grado de correspondencia entre la opinión legal generada y el documento jurídico de referencia. Esta dimensión comprende dos componentes: la similitud léxica, evaluada mediante BLEU, ROUGE-1, ROUGE-2 y ROUGE-L; y la similitud semántica, evaluada mediante BERT Precision, BERT Recall y BERT F1. La consideración conjunta de ambos componentes permite valorar tanto la coincidencia textual como la conservación del significado del contenido jurídico generado.

La dimensión eficiencia corresponde a la capacidad del modelo para generar la opinión legal en un menor tiempo de procesamiento. Esta dimensión se evaluó mediante el tiempo promedio de generación, expresado en segundos, considerando que un menor tiempo representa un mejor desempeño en términos de eficiencia.

En consecuencia, la evaluación de rendimiento quedó definida de la siguiente manera:

Tabla 6 *Calidad y eficiencia demostrados según componentes*

Dimensión	Componentes/indicadores
Calidad	Similitud textual: BLEU, ROUGE-1, ROUGE-2 y ROUGE-L; similitud semántica: BERT Precision, BERT Recall y BERT F1.
Eficiencia	Tiempo promedio de generación.

Nota: Elaboración propia

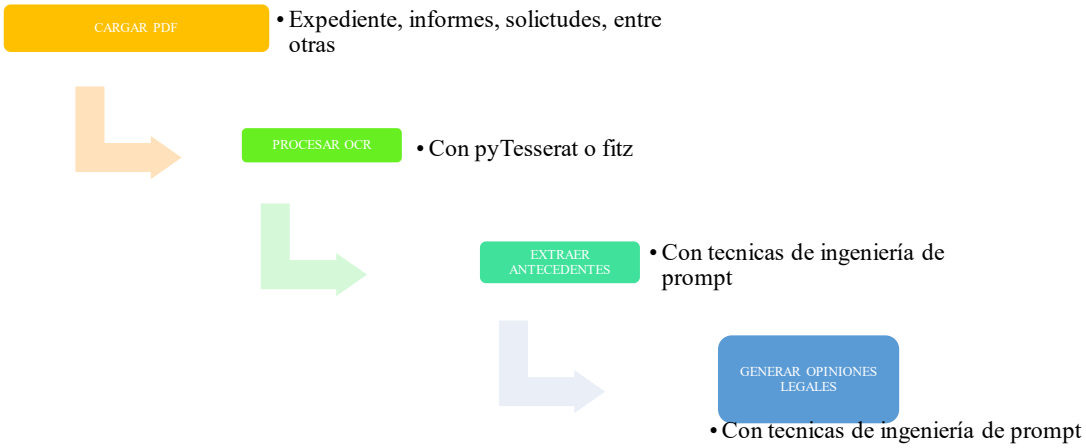
Esta definición permite diferenciar la calidad del contenido generado de la eficiencia del proceso de generación, manteniendo correspondencia con los indicadores y criterios empleados en el análisis multicriterio. La selección de Claude como modelo de mejor

desempeño se realizó considerando conjuntamente ambas dimensiones y los pesos establecidos en la evaluación.

Capitulo V: Propuesta y diseño

5.1 Arquitectura del sistema para mostrar funcionalidad

Figura 9. *Arquitectura del sistema para mostrar la funcionalidad*



Cada etapa representa una fase secuencial del flujo de trabajo desarrollado habitualmente por la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca.

En la siguiente imagen se muestra la interfaz del prototipo funcional desarrollado que utiliza el modelo de lenguaje extenso que tuvo mejor desempeño en el proceso de emisión de opiniones legales.

Figura 10. *Interfaz del sistema que utiliza el abogado*



5.1.1 Carga de documentos PDF

La primera etapa consiste en la carga de documentos digitales en formato PDF, los cuales corresponden a los expedientes administrativos que serán objeto de análisis.

Entre los documentos que pueden ser cargados se encuentran:

- Solicitudes.
- Informes técnicos.
- Memorandos.
- Oficios.
- Resoluciones.
- Contratos.
- Expedientes administrativos.
- Opiniones legales previos.
- Documentación de programas sociales.
- Documentos relacionados con obras públicas y contrataciones.

El sistema permite cargar el documento en formato .pdf, con el objetivo de reunir todos los antecedentes necesarios para la elaboración de la opinión legal.

Asimismo, esta etapa considera la validación del formato de los archivos, verificando que correspondan efectivamente a documentos PDF y que no presenten errores de lectura.

5.1.2 Procesamiento OCR

Una vez cargado el documento, se realiza un proceso de reconocimiento óptico de caracteres (OCR), cuyo propósito es transformar el contenido visual de los archivos PDF en texto legible por el modelo de lenguaje.

Para esta etapa se consideran herramientas como PyTesseract y Fitz (PyMuPDF):

- PyTesseract: permite reconocer texto contenido en imágenes escaneadas.
- Fitz (PyMuPDF): permite extraer texto directamente desde documentos PDF digitales.

El uso de ambas herramientas resulta necesario porque algunos expedientes contienen documentos completamente digitales, mientras que otros corresponden a documentos escaneados.

- Durante esta etapa también se realiza:
- Limpieza de caracteres especiales.
- Eliminación de espacios innecesarios.
- Corrección de saltos de línea.
- Normalización de fechas y numeración.
- Unificación de texto proveniente de múltiples documentos.

El resultado de esta etapa es un texto consolidado y estructurado en formato e texto plano, listo para ser procesado por el modelo de lenguaje extenso (Claude).

5.1.3 Extracción de antecedentes

La tercera etapa consiste en la extracción automatizada de antecedentes relevantes a partir de los documentos procesados (formato texto plano obtenido en la anterior etapa).

Para ello se emplea el modelo Claude Haiku mediante la técnica Plan-and-Solve Prompting.

En la fase de planificación, el modelo identifica:

- Documentos relevantes.
- Fechas importantes.
- Unidades orgánicas o entidades emisoras.

- Remitentes y destinatarios.
- Asuntos principales.
- Hechos relevantes.
- Secuencia cronológica de los acontecimientos.

Posteriormente, en la fase de ejecución, el modelo redacta la sección de antecedentes utilizando un lenguaje claro, formal y ordenado cronológicamente.

Esta etapa permite reducir significativamente el tiempo requerido para revisar manualmente expedientes extensos, especialmente cuando contienen múltiples documentos y antecedentes dispersos.

Asimismo, la técnica Chain of Verification permite verificar que los antecedentes generados:

- Mantengan coherencia cronológica.
- No omitan documentos relevantes.
- No repitan información innecesaria.
- Utilicen una redacción institucional adecuada.

5.1.4 Generación del informe legal

La última etapa consiste en la elaboración automatizada del informe legal a partir de los antecedentes previamente identificados.

En esta fase, Claude Haiku genera:

- Base legal.
- Análisis jurídico.
- Conclusiones.
- Recomendaciones.

- Jurisprudencia aplicable.

La técnica Plan-and-Solve Prompting permite que el modelo estructure previamente el razonamiento jurídico antes de generar la opinión legal final.

En primer lugar, el modelo identifica el tema central del caso, las normas aplicables y la relación entre los antecedentes y el problema jurídico planteado.

Posteriormente, desarrolla el análisis legal considerando:

- Normativa nacional.
- Reglamentos internos.
- Jurisprudencia aplicable.
- Principios administrativos.
- Criterios institucionales.

Finalmente, mediante Chain of Verification, el sistema revisa que:

- Las normas citadas sean coherentes con el caso.
- Las conclusiones respondan al análisis realizado.
- Las recomendaciones sean viables.
- El lenguaje utilizado sea técnico y formal.

5.2 Componentes del sistema

La propuesta del sistema se compone de varios módulos funcionales que interactúan entre sí para permitir el procesamiento automatizado de expedientes y la generación de opiniones legales.

5.2.1 Módulo de carga de archivos

Este módulo permite al usuario seleccionar y cargar documentos PDF desde su equipo.

Su función principal es:

- Validar el formato del archivo.
- Verificar el tamaño del documento.
- Permitir la carga de múltiples archivos.
- Organizar los documentos por expediente.

5.2.2 Módulo OCR

Este módulo transforma el contenido visual de los documentos en texto editable.

Entre sus principales funciones se encuentran:

- Extracción de texto desde PDFs digitales.
- Reconocimiento de texto desde documentos escaneados.
- Corrección de errores básicos de lectura.
- Unificación del contenido textual.

5.2.3 Módulo de extracción de antecedentes

Este módulo utiliza Claude Haiku junto con técnicas de ingeniería de prompts para identificar y redactar automáticamente la sección de antecedentes.

Su función consiste en:

- Detectar fechas y documentos importantes.
- Identificar actores involucrados.
- Organizar los hechos cronológicamente.
- Generar antecedentes resumidos y coherentes.

5.2.4 Módulo de generación del informe legal

Este módulo constituye el núcleo principal del sistema.

Su función consiste en elaborar automáticamente:

- Base legal.

- Análisis jurídico.
- Conclusiones.
- Recomendaciones.

Para ello, utiliza prompts estructurados y mecanismos de verificación que permiten mejorar la calidad del resultado generado.

5.2.5 Módulo de validación

Este módulo tiene como objetivo revisar la coherencia del contenido generado antes de presentar el informe al usuario.

- Entre sus funciones se encuentran:
- Verificar que todas las secciones estén completas.
- Revisar que las normas citadas sean pertinentes.
- Detectar contradicciones entre conclusiones y recomendaciones.
- Evaluar la calidad de la redacción.

5.3 Flujo funcional del sistema

El flujo funcional del sistema inicia con la carga de documentos PDF relacionados con un expediente administrativo. Posteriormente, se ejecuta el proceso OCR para convertir dichos documentos en texto legible. A continuación, el sistema identifica los antecedentes relevantes y organiza cronológicamente la información. Finalmente, el modelo Claude Haiku genera el informe legal completo, incluyendo base legal, análisis, conclusiones y recomendaciones.

Este flujo reproduce de forma simplificada el trabajo habitual realizado por un abogado de la Oficina de Asesoría Jurídica, permitiendo optimizar tiempos de revisión documental y mejorar la consistencia en la elaboración de informes.

5.4 Evaluación humana

La evaluación humana tuvo como finalidad determinar la calidad jurídica de los informes generados por los modelos de lenguaje, mediante la valoración de especialistas con experiencia en la elaboración y revisión de opiniones e informes legales en el ámbito de la administración pública.

La evaluación fue realizada por seis (6) especialistas jurídicos pertenecientes a la Municipalidad Provincial de Calca, quienes cuentan con experiencia en la elaboración y revisión de opiniones legales en materia administrativa. La participación de estos especialistas permitió incorporar el criterio profesional de expertos respecto de la fundamentación jurídica, consistencia argumentativa, precisión normativa, estructura, lenguaje y utilidad práctica de los informes generados.

Con la finalidad de reducir posibles sesgos derivados del conocimiento de la procedencia de los documentos, los informes fueron presentados a los evaluadores bajo un esquema de evaluación ciega (blind review). En consecuencia, los especialistas no tuvieron conocimiento del modelo de lenguaje que había generado cada informe, efectuando la valoración exclusivamente sobre la base del contenido de cada documento.

Para la evaluación se utilizó una ficha estructurada de evaluación, previamente establecida, mediante una escala tipo Likert de cinco puntos, donde: 1 = Muy deficiente

2 = Deficiente

3 = Regular

4 = Bueno

5 = Excelente

Los criterios considerados en la evaluación fueron los siguientes:

- Coherencia de los antecedentes.

- Pertinencia de la base legal.
- Consistencia del análisis jurídico.
- Relación entre conclusiones y recomendaciones.
- Uso adecuado del lenguaje jurídico.
- Estructura formal del informe.
- Precisión normativa.
- Utilidad práctica del documento.
- Nivel de similitud con un informe legal elaborado por un abogado.
- Calidad global del informe generado.

5.4.1 Sustento de los criterios de evaluación

Los criterios empleados en la ficha no fueron establecidos de manera arbitraria, sino que fueron sustentados en referentes jurídico-administrativos, técnicos y profesionales relacionados con la elaboración de documentos jurídicos. Para tal efecto, se consideraron principalmente la Ley N.° 27444, Ley del Procedimiento Administrativo General, la jurisprudencia del Tribunal Constitucional del Perú, el Manual de Técnica Legislativa del Congreso de la República y la ISO 24495-2:2025, Lenguaje claro Parte 2: Comunicación jurídica.

Esta integración permitió evaluar la calidad de los informes desde una perspectiva multidimensional, distinguiendo entre la calidad del contenido jurídico, la calidad de la argumentación, la calidad formal y comunicativa y la utilidad profesional del documento.

5.4.1.1 Ley N.° 27444 y motivación jurídica

La Ley N.° 27444 constituye uno de los principales referentes para la evaluación de la calidad de los informes de naturaleza administrativa. En particular, el artículo 6 establece exigencias relacionadas con la motivación de los actos administrativos, señalando que esta debe

expresar una relación concreta y directa entre los hechos relevantes del caso y las razones jurídicas y normativas que sustentan la decisión. Asimismo, establece que no resultan admisibles fórmulas generales o vacías de fundamentación, ni aquellas que, por su oscuridad, vaguedad, contradicción o insuficiencia, no permitan esclarecer específicamente la motivación del acto.

En este sentido, la coherencia de los antecedentes permite determinar si el informe identifica y desarrolla adecuadamente los hechos relevantes que sirven de base al análisis jurídico. La pertinencia de la base legal permite valorar si las normas citadas guardan correspondencia con el problema jurídico planteado. Por su parte, la consistencia del análisis jurídico permite determinar si existe una relación lógica entre los hechos, las normas aplicables, la interpretación jurídica y la conclusión propuesta.

Asimismo, la relación entre conclusiones y recomendaciones permite evaluar la congruencia interna del documento, verificando que las recomendaciones formuladas sean consecuencia razonada del análisis desarrollado y respondan efectivamente al problema jurídico planteado.

De esta manera, los criterios señalados permiten trasladar a la evaluación empírica las exigencias de motivación previstas en la normativa administrativa, sin limitar la valoración a la mera presencia de citas legales.

5.4.1.2 Jurisprudencia del Tribunal Constitucional

La jurisprudencia del Tribunal Constitucional constituye un segundo referente para sustentar la evaluación de la calidad de los informes jurídicos administrativos, particularmente respecto de la motivación, razonabilidad y ausencia de arbitrariedad.

El Tribunal Constitucional ha establecido que la motivación de la actuación administrativa constituye una exigencia vinculada con el principio de legalidad y que la motivación suficiente constituye una garantía frente a la razonabilidad y la arbitrariedad de las decisiones administrativas. Asimismo, ha señalado que la falta de motivación o su insuficiencia puede constituir una actuación arbitraria e ilegal y resultar contraria a las garantías del debido procedimiento administrativo.

Este criterio resulta relevante para la presente investigación debido a que la evaluación de los informes generados por modelos de lenguaje no puede reducirse a verificar si estos contienen referencias normativas. Es necesario determinar si las normas citadas son pertinentes, si fueron correctamente relacionadas con los hechos y si el razonamiento desarrollado permite justificar las conclusiones alcanzadas.

En consecuencia, la pertinencia de la base legal, la consistencia del análisis jurídico, la precisión normativa y la relación entre conclusiones y recomendaciones constituyen criterios directamente relacionados con la exigencia de una fundamentación jurídica racional y suficiente.

5.4.1.3 Manual de Técnica Legislativa del Congreso de la República

Como referente complementario se consideró el Manual de Técnica Legislativa del Congreso de la República del Perú, en su tercera edición publicada en 2021. El Manual señala que la técnica legislativa comprende, entre otros aspectos, la estructura y redacción de los textos jurídicos, y desarrolla criterios relacionados con el lenguaje, la precisión terminológica, la coherencia, la claridad, la sencillez, la concisión y la precisión.

Si bien el Manual tiene como finalidad principal orientar la elaboración y redacción de textos legislativos, su propio contenido reconoce un ámbito de aplicación más amplio,

señalando que sus criterios pueden ser utilizados, en lo pertinente, para disposiciones de menor jerarquía, incluyendo decretos, resoluciones y ordenanzas municipales. (Congreso del Perú)

Por esta razón, en la presente investigación el Manual no se empleó como una norma que determine directamente la validez de los informes legales evaluados, sino como referente técnico para valorar las características formales y lingüísticas de la redacción jurídica.

Desde esta perspectiva, el criterio uso adecuado del lenguaje jurídico permite valorar la precisión y adecuación de la terminología empleada; mientras que el criterio estructura formal del informe permite valorar la organización lógica y ordenada del documento.

Asimismo, el criterio de precisión normativa se relaciona con la necesidad de emplear correctamente las referencias jurídicas, evitando denominaciones imprecisas o formulaciones que puedan generar ambigüedad respecto de las normas aplicables.

El propio Manual destaca la importancia de la claridad, la coherencia y la precisión para que los textos jurídicos reflejen adecuadamente su contenido y puedan ser comprendidos por sus destinatarios (Congreso del Perú).

5.4.1.4 ISO 24495-2:2025 - Comunicación jurídica

La evaluación también incorporó como referente técnico la ISO 24495-2:2025, Plain language - Part 2: Legal communication, publicada oficialmente por la Organización Internacional de Normalización (ISO) en agosto de 2025. Esta norma está específicamente dirigida a la comunicación jurídica y establece directrices y técnicas destinadas a facilitar que los lectores comprendan la información jurídica y puedan ejercer sus derechos y obligaciones. Asimismo, contempla la necesidad de atender a diferentes destinatarios, respetar los requisitos estructurales y de diseño de los documentos jurídicos, explicar conceptos jurídicos complejos y facilitar la comprensión de los procesos que deben seguir los usuarios.

La ISO 24495-2:2025 establece cuatro principios orientadores de la comunicación jurídica: que los lectores obtengan la información que necesitan, encuentren fácilmente lo que necesitan, comprendan fácilmente lo que encuentran y puedan utilizar fácilmente dicha información.

En la presente investigación, esta norma se utilizó como referente técnico complementario, principalmente para sustentar los criterios relacionados con el uso adecuado del lenguaje jurídico, la estructura formal del informe y la utilidad práctica del documento.

Es importante precisar que la ISO 24495-2:2025 no fue utilizada como una norma para determinar la corrección o validez jurídica del contenido de los informes, sino como un referente especializado para evaluar la calidad de la comunicación jurídica escrita. La corrección jurídica y la fundamentación normativa fueron evaluadas principalmente mediante los criterios derivados de la Ley N.° 27444, la jurisprudencia constitucional y el juicio especializado de los evaluadores.

Esta delimitación resulta importante, puesto que permite evitar una confusión entre calidad jurídica y claridad lingüística: un documento puede estar correctamente redactado y, sin embargo, contener una fundamentación jurídica incorrecta; del mismo modo, puede contener un razonamiento jurídicamente pertinente, pero estar redactado de manera poco clara. Por ello, ambas dimensiones fueron consideradas de manera complementaria.

5.4.1.5 Integración de los criterios de evaluación

A partir de los referentes señalados, los diez criterios incluidos en la ficha de evaluación se sustentaron de la siguiente manera

***Tabla** 7 Integración de criterios de evaluación*

N.°	Criterio de la ficha	Principal sustento
1	Coherencia de los antecedentes	Ley N.° 27444, art. 6 + motivación
2	Pertinencia de la base legal	Ley N.° 27444 + jurisprudencia del TC
3	Consistencia del análisis jurídico	Ley N.° 27444 + jurisprudencia del TC
4	Relación entre conclusiones y recomendaciones	Motivación y congruencia
5	Uso adecuado del lenguaje jurídico	Manual de Técnica Legislativa + ISO 24495-2:2025
6	Estructura formal del informe	Manual de Técnica Legislativa + ISO 24495-2:2025
7	Precisión normativa	Manual de Técnica Legislativa + Ley N.° 27444
8	Utilidad práctica del documento	ISO 24495-2:2025 + juicio experto
9	Nivel de similitud con un informe legal elaborado por un abogado	Juicio experto profesional
10	Calidad global del informe generado	Síntesis del conjunto de criterios

Nota: Elaboración propia

Esta correspondencia permitió mantener sin modificaciones la ficha de evaluación originalmente diseñada, pero otorgando a cada criterio un sustento teórico, jurídico o técnico identificable.

En particular, los primeros cuatro criterios se orientaron principalmente a valorar la calidad del contenido y razonamiento jurídico; los criterios quinto, sexto y séptimo estuvieron dirigidos a valorar la calidad formal, lingüística y normativa del documento; el octavo permitió determinar su utilidad práctica; el noveno permitió contrastar el resultado generado por los modelos de lenguaje con el estándar profesional percibido por los especialistas; y el décimo permitió obtener una valoración global del documento.

5.5 Tiempo total de atención de opiniones legales mediante el método tradicional

Antes de evaluar el impacto del prototipo basado en inteligencia artificial, se estableció una línea base del proceso tradicional de atención de opiniones legales. Para ello, se analizó

una muestra de los expedientes correspondientes a las opiniones legales emitidas por la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca durante el periodo 2022–2024.

El indicador obtenido corresponde al tiempo total transcurrido entre la fecha de solicitud de la opinión legal y la fecha de emisión del documento. Este indicador permite caracterizar el comportamiento histórico del proceso institucional; sin embargo, no representa exclusivamente el tiempo efectivo dedicado por el abogado a las actividades de análisis y elaboración, debido a que puede incluir periodos de espera, asignación, disponibilidad del personal y otras actividades administrativas.

5.5.1 Determinación del tamaño de muestra de los expedientes escaneados

La población de estudio estuvo conformada por 453 opiniones legales emitidas por la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca durante el período 2022–2024. Para determinar el tamaño de la muestra se utilizó la fórmula para poblaciones finitas, considerando un nivel de confianza del 95 % un margen de error del 10 % y una proporción esperada de 0.50, valor recomendado cuando no existen estudios previos que permitan estimar la variabilidad de la población.

La fórmula empleada fue:

$$n = \frac{NxZ^2x\ p\ x\ q}{e^2(N-1) + Z^2x\ p\ x\ q}$$

Sustituyendo los valores:

$$n = \frac{453\ x\ (1.96)^2(0.50)(0.50)}{(0.10)^2(453-1) + (1.96)^2(0.50)\ (0.50)}$$

$$n = 79.39$$

Como el tamaño de muestra debe expresarse en un numero entero, se redondea al entero superior:

n=80 expedientes

Donde:

(N = 453): población de opiniones legales.

(Z = 1.96): valor correspondiente al 95 % de confianza.

(p = 0.50): proporción esperada.

(q = 0.50): proporción complementaria.

(e = 0.10): margen de error permitido.

Al reemplazar los valores en la fórmula se obtuvo un tamaño de muestra de 79.39 expedientes, el cual fue redondeado a 80 expedientes.

En consecuencia, la investigación analizó 80 opiniones legales, seleccionadas mediante muestreo aleatorio simple, garantizando que todos los expedientes de la población tuvieran la misma probabilidad de ser seleccionados. Esta muestra permitió estimar de manera representativa el tiempo promedio de emisión de opiniones legales y realizar el análisis comparativo del proceso tradicional frente al proceso asistido por inteligencia artificial.

5.5.2 Resultados del tiempo de emisión de opiniones legales

Con el propósito de establecer la línea base del proceso tradicional de emisión de opiniones legales, se analizó una muestra de 80 expedientes, seleccionados mediante muestreo aleatorio simple de una población de 453 opiniones legales emitidas por la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca durante el período 2022–2024.

Tabla 8 *Tiempo de emisión de opiniones legales mediante proceso manual*

Expedientes	Fecha de Solicitud de Opinion	Fecha de Emision de Opinion	Dias Demorados
Expediente 1	14/02/2022	21/02/2022	7
Expediente 2	17/02/2022	21/02/2022	4
Expediente 3	3/03/2022	3/03/2022	0
Expediente 4	14/02/2022	25/02/2022	11
Expediente 5	27/01/2022	27/01/2022	0
Expediente 6	26/01/2022	31/01/2022	5
Expediente 7	2/02/2022	2/02/2022	0
Expediente 8	7/02/2022	7/02/2022	0
Expediente 9	4/02/2022	9/02/2022	5
Expediente 10	21/03/2022	23/03/2022	2
Expediente 11	16/03/2022	17/03/2022	1
Expediente 12	10/03/2022	17/03/2022	7
Expediente 13	17/03/2022	25/03/2022	8
Expediente 14	31/03/2022	31/03/2022	0
Expediente 15	24/03/2022	29/03/2022	5
Expediente 16	14/02/2022	15/02/2022	1
Expediente 17	22/03/2022	22/03/2022	0
Expediente 18	22/02/2022	8/03/2022	14
Expediente 19	14/03/2022	14/03/2022	0
Expediente 20	14/03/2022	14/03/2022	0
Expediente 21	14/03/2022	14/03/2022	0
Expediente 22	14/03/2022	14/03/2022	0
Expediente 23	14/02/2022	28/02/2022	14
Expediente 24	14/02/2022	28/02/2022	14
Expediente 25	14/02/2022	28/02/2022	14
Expediente 26	28/02/2022	1/03/2022	1
Expediente 27	14/02/2022	28/02/2022	14
Expediente 28	15/02/2022	1/03/2022	14
Expediente 29	1/03/2022	3/03/2022	2
Expediente 30	30/10/2022	31/10/2022	1
Expediente 31	17/11/2022	18/11/2022	1
Expediente 32	5/12/2022	5/12/2022	0
Expediente 33	26/01/2022	26/01/2022	0
Expediente 34	21/01/2022	24/01/2022	3
Expediente 35	6/01/2022	10/01/2022	4
Expediente 36	14/02/2022	21/02/2022	7
Expediente 37	17/02/2022	21/02/2022	4
Expediente 38	3/03/2022	3/03/2022	0
Expediente 39	14/02/2022	25/02/2022	11
Expediente 40	27/01/2022	27/01/2022	0

Expedientes	Fecha de Solicitud de Opinion	Fecha de Emision de Opinion	Dias Demorados
Expediente 41	26/01/2022	31/01/2022	5
Expediente 42	26/01/2022	31/01/2022	5
Expediente 43	2/02/2022	2/02/2022	0
Expediente 44	7/02/2022	7/02/2022	0
Expediente 45	4/02/2022	9/02/2022	5
Expediente 46	21/03/2022	23/03/2022	2
Expediente 47	16/03/2022	17/03/2022	1
Expediente 48	10/03/2022	17/03/2022	7
Expediente 49	17/03/2022	25/03/2022	8
Expediente 50	31/03/2022	31/03/2022	0
Expediente 51	24/03/2022	29/03/2022	5
Expediente 52	4/04/2022	5/04/2022	1
Expediente 53	22/03/2022	22/03/2022	0
Expediente 54	22/02/2022	8/03/2022	14
Expediente 55	14/03/2022	14/03/2022	0
Expediente 56	14/03/2022	14/03/2022	0
Expediente 57	14/03/2022	14/03/2022	0
Expediente 58	14/02/2022	28/02/2022	14
Expediente 59	14/02/2022	28/02/2022	14
Expediente 60	28/02/2022	1/03/2022	1
Expediente 61	14/02/2022	28/02/2022	14
Expediente 62	28/02/2022	1/03/2022	1
Expediente 63	14/02/2022	28/02/2022	14
Expediente 64	15/02/2022	1/03/2022	14
Expediente 65	1/03/2022	3/03/2022	2
Expediente 66	2/06/2022	2/06/2022	0
Expediente 67	1/06/2022	2/06/2022	1
Expediente 68	6/06/2022	8/06/2022	2
Expediente 69	14/06/2022	15/06/2022	1
Expediente 70	14/06/2022	16/06/2022	2
Expediente 71	22/06/2022	22/06/2022	0
Expediente 72	27/06/2022	27/06/2022	0
Expediente 73	28/06/2022	28/06/2022	0
Expediente 74	28/06/2022	30/06/2022	2
Expediente 75	28/06/2022	1/07/2022	3
Expediente 76	30/06/2022	1/07/2022	1
Expediente 77	28/06/2022	1/07/2022	3
Expediente 78	28/06/2022	4/07/2022	6
Expediente 79	5/07/2022	6/07/2022	1
Expediente 80	4/07/2022	6/07/2022	2

Nota: Elaboración propia

Para cada expediente se registró la fecha de recepción de la solicitud y la fecha de emisión de la opinión legal, determinándose el número de días transcurridos entre ambos, así mismo en la siguiente tabla se detalla la aplicación estadística:

Tabla 9 *Estadísticos descriptivos de tiempo de emisión de opiniones legales mediante procesos manual*

Estadístico	Resultado	Unidad
Población	453	opiniones legales
Muestra	80	Expedientes
Tipo de muestreo	Aleatorio simple	—
Media	4.06	días calendario
Desviación estándar	4.93	Días
Coeficiente de variación	121.48	%
Intervalo de confianza (95 %)	2.96 – 5.16	Días
Mediana	2	Días
Mínimo	0	Días
Máximo	14	Días

Nota: Elaboración propia

Los resultados muestran que el tiempo total de atención institucional presentó una media de 4.06 días calendario, con una desviación estándar de 4.93 días. Esta diferencia evidencia una considerable variabilidad en la duración del proceso, observándose expedientes cuya opinión fue emitida el mismo día de la solicitud y otros que alcanzaron hasta 14 días.

Asimismo, la mediana fue de 2 días, lo que indica que el 50 % de los expedientes presentó un tiempo total de atención igual o inferior a dicho valor. La diferencia entre la media y la mediana evidencia una distribución con mayor concentración de casos en tiempos relativamente cortos y la presencia de expedientes con mayores tiempos de atención.

El coeficiente de variación de 121.48 % evidencia una elevada dispersión relativa de los tiempos observados. Esta variabilidad puede estar asociada a las características particulares de los expedientes, su complejidad, la cantidad de antecedentes que requieren revisión, la

necesidad de identificar normativa aplicable y las condiciones operativas existentes durante el trámite.

En consecuencia, el valor de 4.06 días calendario se establece como indicador histórico del tiempo total de atención del proceso tradicional, y no como una medida directa del tiempo efectivo dedicado por el profesional jurídico a la elaboración de la opinión.

Esta precisión resulta importante para la comparación posterior, debido a que el tiempo total de atención puede incorporar periodos que no corresponden directamente a actividades de procesamiento jurídico.

5.6 Tiempo de elaboración de opiniones legales mediante el prototipo basado en inteligencia artificial

Para evaluar el comportamiento temporal del prototipo desarrollado, se realizó una ejecución controlada del sistema utilizando un expediente administrativo como caso de prueba. El proceso comprendió las etapas de procesamiento inicial del expediente, reconocimiento óptico de caracteres (OCR), extracción de antecedentes relevantes, generación asistida de la opinión legal y revisión o validación por parte del especialista jurídico.

El tiempo registrado para la ejecución completa del flujo fue de aproximadamente 10 minutos, considerando desde el inicio del procesamiento del expediente hasta la finalización de la revisión de la opinión generada.

Este resultado representa el tiempo de ejecución del flujo asistido por inteligencia artificial bajo las condiciones de prueba realizadas. Debido a que la medición histórica del proceso tradicional se obtuvo a partir de fechas administrativas de solicitud y emisión, este valor no debe interpretarse inicialmente como equivalente directo a los 4.06 días, sino como una medición del tiempo de procesamiento del prototipo.

5.6.1 Etapas del proceso asistido mediante inteligencia artificial

El flujo evaluado mediante el prototipo estuvo conformado por las siguientes etapas:

Tabla 10 *Etapas del proceso de elaboración de opiniones legales*

Etapa	Método tradicional	Prototipo basado en IA
Escaneo del expediente	Manual	Asistido por el sistema
Procesamiento OCR	Revisión manual	Extracción asistida
Extracción de antecedentes	Revisión manual	Asistida mediante IA
Elaboración de opinión legal	Redacción manual	Generación asistida por IA
Revisión jurídica	Especialista jurídico	Especialista jurídico
Validación final	Especialista jurídico	Especialista jurídico

Nota: Elaboración propia

La principal diferencia entre ambos métodos se encuentra en las actividades de procesamiento y elaboración. En el método tradicional, estas actividades dependen principalmente de la intervención manual del profesional, mientras que el prototipo automatiza o asiste las tareas de procesamiento documental, extracción de información y generación inicial del contenido.

La validación jurídica permanece a cargo del especialista en ambos métodos, debido a que el propósito del prototipo es brindar apoyo al profesional y no sustituir su responsabilidad en la revisión de la opinión legal.

5.7 Criterio para la comparación del tiempo de elaboración

Para evaluar el comportamiento temporal del proceso de emisión de opiniones legales, se consideraron dos indicadores diferenciados: el tiempo total de atención institucional del proceso tradicional y el tiempo de ejecución del prototipo basado en inteligencia artificial.

Esta diferenciación es necesaria debido a que los registros históricos disponibles únicamente permiten determinar el periodo transcurrido entre la fecha de solicitud y la fecha de

emisión de la opinión legal. Por tanto, no se dispone de información que permita determinar el tiempo efectivo dedicado exclusivamente por el abogado a la elaboración de cada opinión.

5.7.1 Proceso tradicional

A partir de la muestra de 80 expedientes se determinó un tiempo total de atención institucional promedio de 4.06 días calendario, calculado desde la fecha de solicitud hasta la fecha de emisión de la opinión legal.

Este indicador representa el comportamiento histórico del proceso institucional y constituye la línea base temporal de referencia de la investigación.

Sin embargo, este valor no representa exclusivamente el tiempo de trabajo del abogado, debido a que el periodo registrado puede incluir tiempos de espera, asignación, coordinación y otras actividades administrativas.

5.7.2 Prototipo basado en inteligencia artificial

Mediante la ejecución del prototipo se registró un tiempo aproximado de 10 minutos para completar el flujo de procesamiento del expediente, generación de la opinión legal y revisión por parte del especialista jurídico.

El prototipo permite asistir actividades como el procesamiento documental, extracción de antecedentes y generación inicial de la opinión legal, manteniendo la participación del profesional jurídico en la revisión y validación final.

5.7.3 Comparación de los indicadores temporales

Tabla 11. *Comparación temporal del proceso tradicional y el prototipo*

Indicador	Proceso tradicional	Prototipo con IA
Indicador medido	Tiempo total de atención institucional	Tiempo de ejecución del prototipo
Punto de inicio	Fecha de solicitud	Inicio del procesamiento
Punto de finalización	Fecha de emisión	Generación y validación
Resultado	**4.06 días calendario**	**≈ 10 minutos**
Naturaleza del indicador	Histórico	Experimental

Nota: Elaboración propia

Los resultados evidencian que el proceso tradicional presenta un tiempo total de atención institucional promedio de 4.06 días calendario, mientras que el prototipo puede completar su flujo de procesamiento en aproximadamente 10 minutos. No obstante, estos valores no se consideran directamente comparables para determinar un porcentaje de reducción del tiempo, debido a que fueron obtenidos bajo diferentes puntos de inicio y finalización.

Por tanto, el resultado permite evidenciar el potencial de reducción temporal asociado a la automatización y asistencia mediante inteligencia artificial, pero no permite afirmar estadísticamente que el prototipo reduzca en un determinado porcentaje los 4.06 días del proceso institucional.

5.7.4 Consideración metodológica

La principal limitación para establecer una comparación directa corresponde a la inexistencia de registros históricos sobre el tiempo efectivo de trabajo del abogado. Los documentos disponibles registran fechas de solicitud y emisión, pero no permiten identificar

cuánto tiempo fue dedicado específicamente al análisis, búsqueda normativa, redacción y revisión de cada opinión legal.

En consecuencia, los 4.06 días se mantienen como indicador de tiempo total de atención institucional**,** mientras que los 10 minutos corresponden al tiempo observado durante la ejecución del prototipo. Para establecer en trabajos posteriores una comparación experimental directa, será necesario registrar el tiempo efectivo dedicado a la elaboración de una misma muestra de expedientes mediante ambos métodos.

Capítulo VI: Resultados y Análisis

En el presente capítulo se exponen y analizan de manera sistemática los resultados obtenidos durante el desarrollo e implementación del prototipo funcional basado en inteligencia artificial generativa, orientado a la automatización del proceso de emisión de opiniones legales en la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca. Asimismo, se evalúa el desempeño del sistema en términos de eficiencia, precisión y aporte a la mejora de la gestión jurídica institucional.

La estructura del capítulo se desarrolla en función de los objetivos específicos planteados en la investigación. En primer lugar, se expone la evaluación comparativa realizada entre distintos Modelos de lenguaje a gran escala (LLM), con la finalidad de identificar la alternativa con mejor desempeño para la generación automatizada de opiniones legales. Posteriormente, se describe el desarrollo e implementación del prototipo funcional basado en el modelo seleccionado. Finalmente, se presentan los resultados de la evaluación de calidad de las opiniones legales generadas por el sistema, mediante la aplicación de instrumentos validados por especialistas jurídicos.

Para el desarrollo experimental se consideraron los modelos Claude, Llama y Mistral. En el caso del modelo Claude, se aplicaron técnicas avanzadas de ingeniería de prompts, específicamente Plan and Solve Prompting y Chain of Verification Prompting, mientras que los modelos Llama y Mistral fueron sometidos a procesos de ajuste mediante fine-tuning utilizando el dataset jurídico construido para la investigación.

Los resultados fueron obtenidos a partir de evaluaciones cuantitativas realizadas mediante una ficha de evaluación validada por especialistas jurídicos, utilizando una escala Likert de 1 a 5 puntos, donde 1 representa un desempeño totalmente deficiente y 5 un desempeño excelente. Los criterios considerados incluyeron la coherencia jurídica, la

adecuación normativa, la precisión en la interpretación de antecedentes, la estructura formal, la claridad de redacción, la fundamentación argumentativa, la utilidad práctica y el nivel de autonomía del sistema.

6.1 Resultados correspondientes al objetivo específico 1: Comparación de Modelos de lenguaje a gran escala

Con la finalidad de identificar el modelo de lenguaje a gran escala con mejor desempeño para la generación automatizada de opiniones legales, se realizó una evaluación comparativa de los modelos Claude, Llama 3.1 y Mistral, considerando tanto métricas de similitud léxica y semántica como indicadores de eficiencia de generación.

La comparación se realizó utilizando el dataset jurídico especializado construido para la investigación, compuesto por 453 documentos correspondientes a opiniones legales emitidas por la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca durante el periodo 2022–2024. Para garantizar condiciones homogéneas de evaluación, los modelos fueron evaluados utilizando documentos pertenecientes al conjunto establecido para la comparación y bajo criterios equivalentes de generación.

En el caso de Llama 3.1 y Mistral, se realizó un proceso de fine-tuning utilizando el dataset jurídico estructurado en formato instrucción – entrada - salida. Por su parte, Claude fue evaluado mediante técnicas de ingeniería de prompts, específicamente Plan and Solve Prompting y Chain of Verification, debido a que el modelo no fue sometido a un proceso de ajuste de parámetros dentro de la presente investigación. Esta diferencia responde a las características de acceso y configuración de cada modelo, por lo que la comparación se orientó a determinar qué configuración modelo ajustado mediante fine-tuning o modelo utilizado mediante técnicas avanzadas de prompting ofrecía mejores resultados para el contexto específico de generación de opiniones legales.

Los resultados de la Tabla 3 muestran diferencias en el desempeño de los tres modelos. En cuanto a la similitud léxica, medida mediante BLEU, el modelo Llama 3.1 obtuvo el valor más alto (0.091652), seguido de Claude (0.058738) y Mistral (0.036796). Este resultado indica que Llama presentó una mayor coincidencia de secuencias de palabras con respecto a los documentos de referencia.

De manera similar, Llama 3.1 alcanzó el mayor valor de ROUGE-1, con 0.294274, y de ROUGE-L, con 0.173982, superando ligeramente a Claude, que obtuvo valores de 0.287010 y 0.169072, respectivamente. Mistral presentó los valores más bajos en ambas métricas, con 0.159672 para ROUGE-1 y 0.106364 para ROUGE-L. Estos resultados evidencian que Llama logró una mayor coincidencia superficial y estructural con los textos de referencia en determinadas dimensiones.

Sin embargo, al considerar ROUGE-2, Claude presentó el mejor resultado, con un valor de 0.201403, frente a 0.134796 obtenido por Llama 3.1 y 0.065791 por Mistral. Este indicador permite observar coincidencias de secuencias de dos palabras y, en este caso, evidencia una mayor correspondencia entre el contenido generado por Claude y los documentos de referencia.

Una diferencia más significativa se observa en las métricas de BERTScore, orientadas a evaluar la similitud semántica entre el texto generado y el documento de referencia. Claude obtuvo una precisión de 0.739016, un recall de 0.735489 y un F1 de 0.737248, superando a Llama 3.1, que obtuvo 0.703614, 0.670019 y 0.685780, respectivamente. Mistral presentó los valores más bajos, con una precisión de 0.294822, un recall de 0.588669 y un F1 de 0.606974. En consecuencia, desde la perspectiva de similitud semántica, Claude presentó el desempeño más favorable de los tres modelos evaluados.

Respecto al Exact Match, los tres modelos obtuvieron un valor de 0.000000. Este resultado indica que ninguna de las generaciones coincidió de manera idéntica con el

documento de referencia. No obstante, este indicador debe interpretarse con cautela en el contexto de la presente investigación, debido a que las opiniones legales son documentos extensos y estructurados, y una generación puede expresar contenidos jurídicamente equivalentes utilizando una redacción diferente. Por ello, el Exact Match fue considerado como un indicador complementario y no como criterio único para determinar el desempeño de los modelos.

En relación con la eficiencia de generación, Claude presentó un tiempo promedio de inferencia de 33.814 segundos, considerablemente inferior al registrado por Llama 3.1 (105.961 segundos) y Mistral (114.086 segundos). Asimismo, Claude generó un promedio de 4.000 tokens, mientras que Llama 3.1 y Mistral generaron 1.673,582 y 1.831,066 tokens, respectivamente. Este resultado evidencia que Claude produjo una mayor cantidad de contenido en un menor tiempo promedio de generación, constituyendo una ventaja relevante para su utilización dentro del prototipo funcional.

En términos generales, los resultados muestran que Llama 3.1 presentó ventajas en las métricas BLEU, ROUGE-1 y ROUGE-L, lo que evidencia una mayor coincidencia léxica y estructural con los documentos de referencia. Sin embargo, Claude obtuvo los mejores resultados en ROUGE-2 y en las tres métricas de BERTScore, además de presentar el menor tiempo promedio de inferencia y la mayor cantidad de tokens generados. Por su parte, Mistral presentó los valores más bajos en la mayoría de las métricas de similitud y el mayor tiempo promedio de generación.

Por tanto, la selección del modelo para el prototipo funcional no se realizó considerando una única métrica, sino mediante una evaluación integral del desempeño textual, semántico y computacional. Bajo estos criterios, Claude fue seleccionado como la alternativa más adecuada para la implementación del prototipo, debido principalmente a su mayor similitud semántica

medida mediante BERTScore, su desempeño en ROUGE-2 y su mayor eficiencia temporal de generación.

Estos resultados deben interpretarse considerando que los modelos fueron evaluados bajo configuraciones diferentes: Llama 3.1 y Mistral mediante fine-tuning, mientras que Claude mediante técnicas de ingeniería de prompts. En consecuencia, los resultados no permiten afirmar que una técnica sea universalmente superior a la otra, sino que permiten determinar qué combinación de modelo y estrategia de adaptación presentó el comportamiento más favorable para el problema específico abordado en esta investigación.

Los resultados obtenidos guardan relación con los hallazgos reportados por Fuentes de Abreu (2025) en su investigación Generación de Resúmenes Judiciales utilizando Modelos de Lenguaje (LLMs), en la cual se evaluó el uso de modelos de lenguaje para la generación automatizada de textos jurídicos. Según el estudio citado, Claude presentó resultados favorables en aspectos relacionados con la coherencia, fidelidad y precisión de los textos generados. Esta coincidencia permite considerar que el desempeño observado en la presente investigación es consistente con antecedentes que evidencian la utilidad de los modelos de lenguaje para la generación de contenido jurídico.

Finalmente, a partir de los resultados cuantitativos obtenidos, Claude fue seleccionado para la implementación del prototipo funcional, debido a que presentó el mejor comportamiento global al combinar similitud semántica, correspondencia textual y eficiencia de generación. Esta selección permitió continuar con la siguiente etapa de la investigación, correspondiente al desarrollo y evaluación del prototipo funcional para la generación automatizada de opiniones legales.

6.2 Resultados correspondientes al objetivo específico 2: Desarrollo del prototipo funcional

Una vez identificado el modelo de lenguaje a gran escala con mejor desempeño en la fase comparativa, se procedió al desarrollo e implementación de un prototipo funcional basado en inteligencia artificial generativa, orientado a la automatización del procesamiento de expedientes administrativos y la generación de opiniones legales en la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca.

El prototipo fue desarrollado utilizando el modelo Claude, debido a que presentó los mejores resultados en criterios de coherencia jurídica, adecuación normativa, claridad argumentativa y estructura formal durante la evaluación comparativa realizada previamente.

El funcionamiento del sistema fue diseñado bajo un flujo de procesamiento secuencial, iniciando con el ingreso del expediente administrativo por parte del usuario. Posteriormente, el sistema procesa la información contenida en los documentos administrativos, identifica los antecedentes relevantes y genera automáticamente una propuesta de opinión legal utilizando técnicas avanzadas de ingeniería de prompts.

Para mejorar la calidad de los resultados generados, se implementaron las técnicas de ingeniería de prompts (Plan and Solve Prompting y Chain of Verification Prompting). La técnica Plan and Solve permitió estructurar el razonamiento jurídico de manera secuencial y organizada, favoreciendo una adecuada interpretación de los antecedentes administrativos y de la normativa aplicable. Por su parte, la técnica Chain of Verification incorporó una etapa de verificación interna orientada a reducir inconsistencias, omisiones normativas y errores argumentativos en la respuesta final generada por el modelo.

El prototipo fue diseñado considerando la estructura formal utilizada por la Oficina de Asesoría Jurídica, permitiendo generar documentos con una organización similar a la utilizada

en las opiniones legales institucionales. Asimismo, el sistema fue configurado para mantener un lenguaje técnico-jurídico acorde con el estándar documental de la entidad.

Durante las pruebas funcionales, el sistema demostró capacidad para procesar expedientes administrativos y generar opiniones legales preliminares de manera automatizada, manteniendo coherencia en la estructura argumentativa y adecuada relación entre antecedentes, análisis jurídico y conclusiones.

Los resultados obtenidos evidencian que el prototipo funcional desarrollado permite automatizar tareas relacionadas con la revisión documental y la elaboración preliminar de opiniones legales, constituyendo una herramienta de apoyo para optimizar el trabajo del personal jurídico.

6.3 Resultados correspondientes al objetivo específico 3: Evaluación de la calidad de las opiniones legales generadas por el prototipo funcional

Con la finalidad de evaluar la calidad de las opiniones legales generadas por el prototipo funcional, se realizó una evaluación humana mediante la participación de seis (6) especialistas jurídicos pertenecientes a la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca.

Para la evaluación de la dimensión "Usabilidad e Interacción con el Usuario" del sistema de Inteligencia Artificial, se adoptó un enfoque metodológico simplificado basado en los principios de evaluación heurística desarrollados por Nielsen (1994a). Siguiendo las recomendaciones del experto validador y con el fin de optimizar la aplicabilidad del instrumento en el contexto administrativo municipal, se optó por sintetizar las diez heurísticas tradicionales en dos macro-criterios operativos: Facilidad de operación e interfaz y Nivel de autonomía del sistema.

El primer macro-criterio agrupó los principios de consistencia, reconocimiento y diseño minimalista, asegurando la eficiencia del operador legal; mientras que el segundo midió el control y la libertad del usuario respecto a la corrección del borrador automatizado. Esta estructuración ejecutiva encuentra sustento teórico en Nielsen (1994b) y Jiménez y Ramírez (2020), quienes señalan que en entornos corporativos o gubernamentales de alta carga laboral, las evaluaciones heurísticas deben ser ágiles, adaptadas al flujo de trabajo real y simplificadas para evitar la saturación del instrumento técnico, garantizando de este modo una recolección de datos precisa y de alto valor pragmático (Sauro, 2016).

La evaluación se desarrolló utilizando expedientes administrativos reales como casos de prueba. Cada especialista interactuó directamente con el sistema y revisó las opiniones legales generadas por el prototipo, emitiendo posteriormente una valoración técnica mediante una ficha de evaluación validada para la investigación.

El instrumento utilizado se basó en una escala Likert de 5 niveles, donde:

- 1 muy deficiente,
- 2 deficiente,
- 3 regular,
- 4 bueno,
- 5 excelente.

Los criterios considerados en la evaluación fueron:

- coherencia jurídica,
- adecuación normativa,
- precisión en la interpretación de antecedentes,
- estructura formal,
- claridad jurídica,

- fundamentación argumentativa,
- utilidad práctica,
- nivel de autonomía del sistema.

Los resultados consolidados de la evaluación realizada por los especialistas se presentan en la

Tabla 12 *Evaluación del prototipo por especialistas jurídicos*

Criterio	Evaluador 1	Evaluador 2	Evaluador 3	Evaluador 4	Evaluador 5	Evaluador 6	Promedio
Coherencia jurídica	5	5	4	5	5	4	4.67
Adecuación normativa	5	4	4	5	5	4	4.5
Precisión de antecedentes	4	4	4	5	4	4	4.17
Estructura formal	5	5	4	5	5	5	4.83
Claridad jurídica	5	4	4	5	5	4	4.5
Fundamentación	4	4	4	5	4	4	4.17
Utilidad práctica	5	5	5	5	5	5	5
Nivel de autonomía	4	4	4	5	4	4	4.17
Evaluación global	5	4	4	5	5	4	4.5

Nota: Elaboración propia

De acuerdo con los resultados obtenidos, el sistema alcanzó niveles altos de aceptación en la mayoría de los criterios evaluados. El criterio con mayor puntuación fue utilidad práctica, obteniendo un promedio de 5 puntos, lo que evidencia que los especialistas consideran que el prototipo puede ser utilizado como herramienta de apoyo dentro de la Oficina de Asesoría Jurídica.

Asimismo, los criterios relacionados con coherencia jurídica (4.67), estructura formal (4.83) y adecuación normativa (4.5) obtuvieron valoraciones elevadas, indicando que las

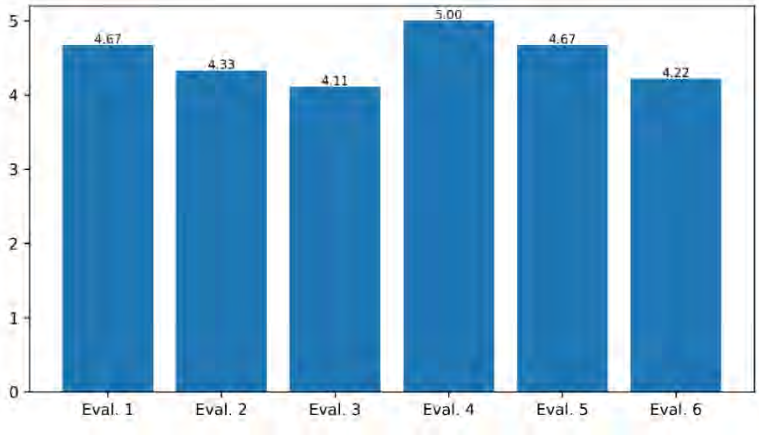
opiniones legales generadas presentan un nivel adecuado de consistencia técnica y organización jurídica.

En relación con la precisión de antecedentes, fundamentación argumentativa y nivel de autonomía, los resultados también mostraron promedios favorables, aunque ligeramente inferiores respecto a otros criterios. Esto evidencia que, si bien el sistema presenta un desempeño sólido, aún requiere supervisión humana en determinados casos complejos o cuando existen escenarios jurídicos con alto nivel de interpretación normativa.

Los resultados permiten afirmar que el prototipo funcional desarrollado posee una alta capacidad para apoyar el proceso de emisión de opiniones legales, contribuyendo a mejorar la rapidez en la elaboración documental y fortaleciendo la eficiencia operativa del área jurídica.

Asimismo, la evaluación realizada por especialistas evidencia que la implementación de inteligencia artificial generativa puede constituir una alternativa viable para optimizar procesos documentales en contextos institucionales, siempre que exista supervisión profesional durante la validación final de los informes generados.

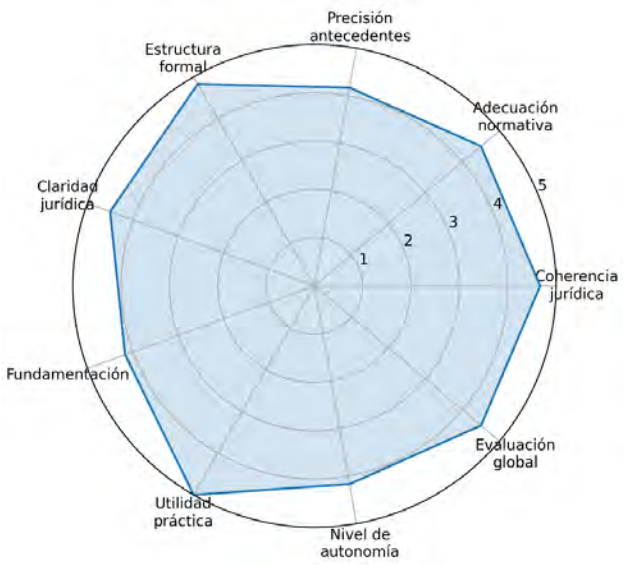
Figura 11. *Gráfico de barras por criterio del prototipo*



La figura presenta el promedio de las calificaciones otorgadas por los seis especialistas jurídicos en cada uno de los criterios de evaluación del prototipo. Se observa que la utilidad

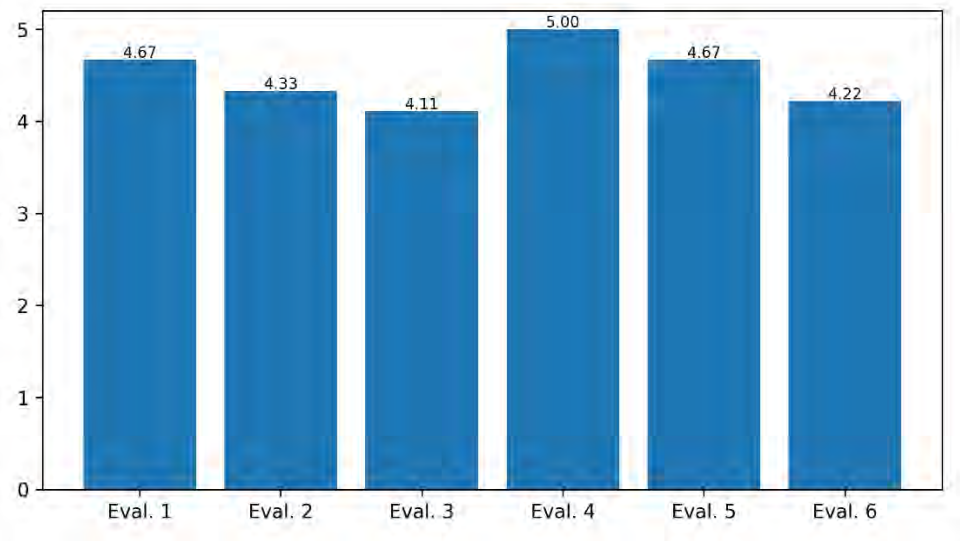
práctica obtuvo la máxima puntuación (5.00), seguida de la estructura formal (4.83) y la coherencia jurídica (4.67), evidenciando un alto nivel de aceptación del sistema en la generación de resoluciones administrativas.

Figura 12. *Gráfico de radar del desempeño del prototipo*



La figura muestra el perfil de desempeño del prototipo en los nueve criterios evaluados. Se aprecia un comportamiento uniforme con calificaciones superiores a 4.0 en todas las dimensiones, destacando la utilidad práctica como el criterio mejor valorado, lo que demuestra que el sistema mantiene un desempeño equilibrado y consistente desde la perspectiva de los especialistas jurídicos.

Figura 13. *Gráfico de barras del promedio por especialista*



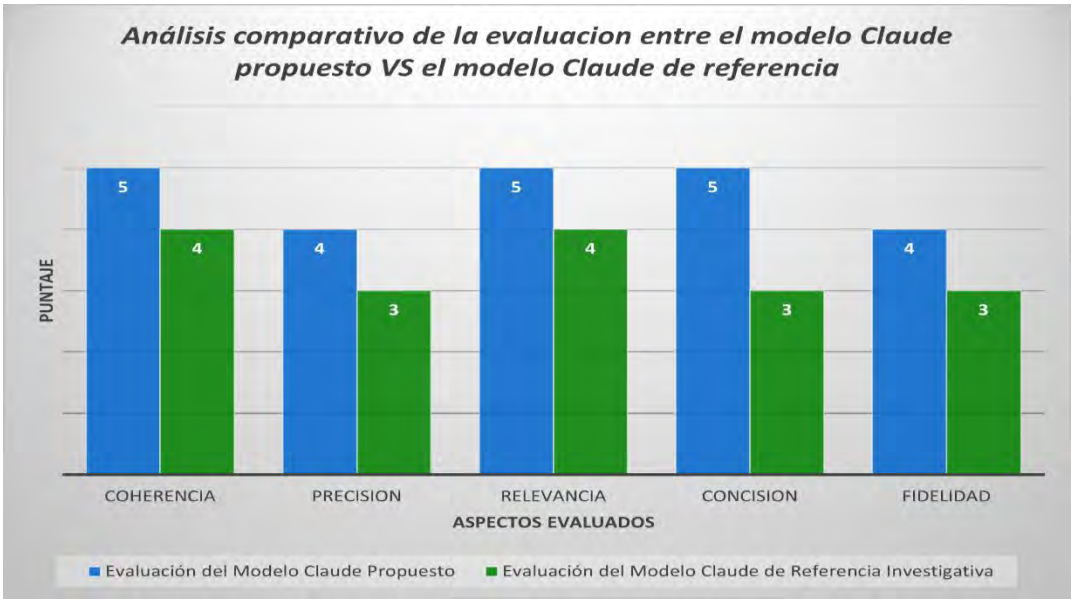
La figura presenta el promedio de las calificaciones otorgadas por cada uno de los seis especialistas jurídicos. Se observa que todos los evaluadores asignaron puntuaciones superiores a 4.0, siendo el Evaluador 4 quien otorgó la máxima calificación (5.00). Estos resultados reflejan un elevado nivel de consenso y aceptación del prototipo entre los expertos participantes.

6.4 Discusión de resultados

Los resultados obtenidos en la presente investigación evidencian que los Modelos de lenguaje a gran escala pueden constituir herramientas viables para apoyar procesos jurídicos relacionados con la elaboración de opiniones legales en entidades públicas.

En relación con el primer objetivo específico, se determinó que el modelo Claude presentó el mejor desempeño general frente a Llama y Mistral, alcanzando mayores niveles de coherencia jurídica, adecuación normativa, claridad argumentativa y utilidad práctica. Estos resultados coinciden con lo reportado por Fuentes de Abreu (2025), quien identificó que Claude generaba resúmenes jurídicos más coherentes y estructurados en comparación con Llama 3.1. En el siguiente diagrama se presenta un análisis comparativo de los puntajes obtenidos por el modelo Claude desarrollado en la presente investigación y los puntajes reportados en el estudio de referencia, Fuentes de Abreu (2025).

Figura 14. *Análisis comparativo de la evaluación entre el modelo Claude propuesto VS el modelo Claude de referencia*



De igual manera, ambos estudios coinciden en que los modelos open-source presentan mayores limitaciones en tareas jurídicas complejas, especialmente en aspectos relacionados con precisión normativa, fidelidad contextual y mantenimiento de estructuras formales consistentes.

Respecto al segundo objetivo específico, relacionado con el desarrollo del prototipo funcional, los resultados obtenidos demuestran que la implementación de inteligencia artificial generativa permitió automatizar parcialmente el procesamiento de expedientes administrativos y la generación preliminar de opiniones legales. En este sentido, la percepción positiva de los especialistas jurídicos evidencia que el sistema posee aplicabilidad práctica dentro del contexto institucional.

Estos hallazgos guardan relación con lo señalado por Fuentes de Abreu (2025), quien sostiene que los sistemas basados en LLM pueden contribuir significativamente a reducir tiempos de trabajo y facilitar la generación de borradores jurídicos preliminares, sin eliminar la necesidad de revisión profesional.

En relación con el tercer objetivo específico, se evidenció que la calidad de las opiniones legales generadas alcanzó niveles altos en criterios de coherencia jurídica, estructura formal y claridad de redacción. Asimismo, se identificó que la aplicación de técnicas de ingeniería de prompts, específicamente Plan and Solve y Chain of Verification, contribuyó significativamente a mejorar la precisión de las respuestas y reducir inconsistencias argumentativas.

Estos resultados coinciden con investigaciones previas que sostienen que la calidad de los resultados generados por los LLM depende en gran medida de la estructura de las instrucciones proporcionadas al modelo.

No obstante, la investigación también permitió identificar limitaciones importantes. Aunque Claude presentó un desempeño superior, aún requiere supervisión humana en casos jurídicos complejos, especialmente cuando existen conflictos interpretativos, vacíos normativos o necesidad de contextualización institucional específica. Esta limitación también fue identificada en la investigación antecedente, donde se concluye que los modelos de lenguaje pueden generar información adicional innecesaria o presentar omisiones relevantes si no existe validación profesional posterior.

En términos generales, los resultados permiten afirmar que la implementación de inteligencia artificial generativa en la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca constituye una alternativa viable para optimizar procesos internos, reducir tiempos de elaboración documental y fortalecer la productividad institucional, manteniendo siempre la supervisión técnica del profesional jurídico.

Conclusiones

Primera. Se determinó que la implementación de Modelos de lenguaje a gran escala (LLM) permite automatizar el proceso de emisión de opiniones legales en la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca, evidenciándose mejoras en la eficiencia operativa, la reducción de tiempos de respuesta y la calidad de los documentos generados.

Segunda. A partir de la evaluación comparativa, se identificó que el modelo Claude presenta el mejor desempeño entre los modelos analizados (Claude, Llama y Mistral), alcanzando el mayor promedio en criterios como coherencia jurídica, adecuación normativa, precisión en la interpretación de antecedentes y estructura formal, lo que lo posiciona como la alternativa más adecuada para su aplicación en la generación de opiniones legales.

Tercera. Se logró desarrollar un prototipo funcional basado en inteligencia artificial generativa, capaz de procesar expedientes administrativos y generar opiniones legales de manera automatizada, integrando técnicas de ingeniería de prompts que permiten estructurar adecuadamente el razonamiento jurídico y mejorar la calidad de salida del sistema.

Cuarta. La evaluación de la calidad de los informes generados evidenció que el sistema alcanza niveles altos en coherencia jurídica, precisión normativa, claridad de redacción y fundamentación argumentativa, de acuerdo con las valoraciones realizadas mediante una ficha de evaluación validada por especialistas jurídicos.

Quinta. Los resultados de la evaluación humana, realizada por abogados especialistas, evidencian que el sistema posee una alta utilidad práctica como herramienta de apoyo en el entorno institucional, contribuyendo a mejorar la productividad y facilitando la revisión de expedientes, aunque manteniendo la necesidad de supervisión profesional en casos complejos.

Recomendaciones

En función de los resultados obtenidos en la presente investigación, se formulan las siguientes recomendaciones orientadas a fortalecer la implementación, optimización y sostenibilidad del sistema basado en inteligencia artificial para la redacción de opiniones legales, así como a maximizar su impacto en la eficiencia de la gestión administrativa dentro de la Municipalidad Provincial de Calca:

Primera. Se recomienda implementar progresivamente el sistema basado en el modelo Claude dentro de la Oficina de Asesoría Jurídica de la Municipalidad Provincial de Calca, inicialmente como una herramienta de apoyo para la redacción preliminar de opiniones legales.

Segunda. Se recomienda complementar el sistema con una base de datos normativa actualizada, que incluya leyes, reglamentos, jurisprudencia, directivas y pronunciamientos administrativos, con la finalidad de incrementar la precisión de las respuestas generadas.

Tercera. Se recomienda capacitar al personal jurídico en el uso de herramientas de inteligencia artificial e ingeniería de prompts, a fin de maximizar el aprovechamiento del sistema y mejorar la interacción entre los abogados y el modelo de lenguaje.

Cuarta. Se recomienda incorporar mecanismos de validación automática y revisión humana antes de la emisión final de cualquier informe legal, con el objetivo de reducir riesgos derivados de errores normativos o interpretaciones inadecuadas.

Quinta. Se recomienda ampliar el alcance del sistema a otras áreas de la Municipalidad Provincial de Calca, como logística, recursos humanos, administración tributaria y gestión documentaria, donde también se generan informes y documentos administrativos de manera recurrente.

Trabajos a futuro

Como trabajo a futuro, se propone evolucionar el prototipo funcional desarrollado con Claude mediante la integración de una arquitectura de Retrieval-Augmented Generation (RAG), que permita recuperar automáticamente la normativa relacionada con cada expediente antes de generar la opinión legal. Para ello, se plantea implementar una base de conocimiento jurídica actualizada que incorpore legislación vigente, reglamentos, jurisprudencia, ordenanzas municipales e informes legales institucionales. La información deberá ser procesada y organizada considerando su vigencia, fecha de publicación, tipo de norma y entidad emisora, permitiendo proporcionar a Claude fuentes normativas como contexto y mostrar las referencias utilizadas para facilitar su verificación por parte del abogado responsable.

Asimismo, se propone evolucionar el prototipo hacia un sistema real para la Oficina de Asesoría Jurídica, incorporando autenticación de usuarios, control de acceso según roles, registro de actividades y trazabilidad de las operaciones realizadas sobre los expedientes. Considerando que el sistema procesaría documentación administrativa que podría contener datos personales, su desarrollo deberá contemplar medidas de seguridad y protección de la información de acuerdo con la Ley N.° 29733, Ley de Protección de Datos Personales, y la normativa aplicable.

Bibliografía

(s.f.).

Arias-Gómez, J., Villasís-Keever, M. Á., & Miranda Novales, M. G. (Eds.). (2016). El protocolo de investigación III: la población de estudio. *El protocolo de investigación III: la población de estudio, 63*. Ciudad de México, México: Revista Alergia.

Ashley, K., & Phelps, T. (2022). "ALEXA,WRITE A MEMO": THE PROMISE AND CHALLENGES OF AIAND LEGAL WRITING. *Revista del Instituto de Redacción Jurídica*, 59.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, . En *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* (págs. 610-623).

Brown, T. M., & Amodei, D. (2020). Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems. NeurIPS.

Devlin, J. C., & Toutanova, K. (2019). ERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL.

Doroteo Guerrero, F. V., & Camacho Ortiz, M. E. (2021). INTELIGENCIA ARTIFICIAL Y LA AUTOMATIZACIÓN DE PROCESOS JUDICIALES EN LA ADMINISTRACIÓN DE JUSTICIA DEL PERÚ. *Tesis para optar el título profesional de Abogada*. Universidad Privada del Norte, Lima.

Escobar, R. S. (12 de agosto de 2022). *Qué es una API y cómo funciona*. Obtenido de Openwebinars.net.: https://openwebinars.net/blog/que-es-una-api-y-como-funciona/

Fernández, T. (2022). *Innovación y Modernización en la Gestión Pública: El Rol de la Automatización.* Editorial Civitas.

García Marín, J. A. (2020). *nteligencia Artificial y Derecho: Automatización de Procesos Legales.* Aranzadi.

García, M., & López, A. (2021). *Inteligencia Artificial y Derecho: Fundamentos y Aplicaciones.* Editorial Tecnos.

García, M., & López, A. (2021). *Marco Jurídico de la Administración Pública: Normativas y Regulaciones Clave.* Editorial Tecnos.

García, M., & Martínez, A. (2021). *Derecho y Tecnología: La Evolución de la Opinión Legal en la Era Digital.* Editorial Tecnos.

García, M., & Pérez, J. (2023). *Inteligencia Artificial y Derecho: Capacitación y Gestión del Cambio.* Editorial Reus.

Hadi, M. U., Al-Tashi, Q., Qureshi, R., Shah, A., Muneer, A., Muhammad, . . . Shah, M. (2023). LLMs: A Comprehensive Survey of Applications, Challenges, Datasets, Limitations, and Future Prospects. *LLMs: A Comprehensive Survey of Applications, Challenges, Datasets, Limitations, and Future Prospects*. arXiv.

Hernández Sampieri, R. ,. (2018). *Metodología de la investigación (6° ed.).* Mexico.

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing (3rd ed.).* Pearson.

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing (3rd ed.).* Pearson.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning. En J. M. Lafferty, & F. C. Pereira,

Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning (págs. (pp. 282-289).).

López, M., & Fernández, J. (2021). *Transformación Digital y Administración Pública: Nuevas Herramientas para la Eficiencia.* Tecnos.

López, M., & Hernández, A. (2022). *Integración de Inteligencia Artificial en Sistemas Jurídicos: El Uso de APIs.* Editorial Tecnos.

López, M., & Martínez, C. (2022). *Capacitación y Transformación Digital en el Ámbito Jurídico: Estrategias para el Siglo XXI.* Editorial Tecnos.

López, M., & Silva, J. (2021). *Transformación Digital en la Administración Pública: El Impacto de la IA.* Editorial Tecnos.

López, R., & García, J. (2023). *Regulación y Modernización del Estado: El Impacto de las Normativas en la Automatización Legal.* Tirant lo Blanch.

López, R., & Hernández, J. (2022). *Interpretación Jurídica y Redacción de Documentos Legales: Guía para Profesionales.* Aranzadi.

Lucidchart.com. (01 de diciembre de 2017). *Las etapas del ciclo de vida de desarrollo de software según la metodología ágil*. Obtenido de https://www.lucidchart.com/blog/es/ciclo-de-vida-del-desarrollo-de-software-segun-la-metodologia-agil

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Martín, P., & Hernández, F. (2020). *Capacitación en Nuevas Tecnologías para la Administración Pública.* Editorial Tirant lo Blanch.

Martín, P., & Hernández, F. (2023). *Automatización en el Sector Público: Casos de Éxito en Europa*. Tirant lo Blanch.

Martínez, C., & Gómez, L. (2021). *Automatización Legal: La Precisión en la Redacción y Revisión de Documentos Jurídicos.* Editorial Reus.

Martínez, C., & González, L. (2022). *Automatización y Derecho: El Impacto de la IA en la Práctica Legal.* Editorial Reus.

Martínez, C., & González, L. (2023). *Automatización y Derecho: El Impacto de la IA en la Práctica Legal.* Editorial Reus.

Martínez, C., & Pérez, L. (2022). *Automatización y Derecho: Normativas Legales en la Era Digital.* Editorial Reus.

Martínez, C., & Pérez, L. (2022). *Inteligencia Artificial y Derecho: Fundamentos del Procesamiento de Lenguaje Natural.* Editorial Tecnos.

Martínez, C., & Pérez, L. (2023). *Automatización en el Derecho: Nuevas Fronteras en la Redacción Legal.* Tirant lo Blanch.

Mongoli, A. (2024). The Use of LLMs in the Legal Field: Optimizing Contract Management with Generative Artificial Intelligence. *Tesis de Maestría.* Politecnico di Torino, ITALIA.

Noe, R. A. (2021). *Employee Training & Development (8th ed.).* McGraw-Hill Education.

Omer, A. (12 de 08 de 2021). *API management is the process of managing different API functions like API creation, publication, securing, and monitoring. While API management tools make it easy for your organization to develop, manage, and secure APIs, they offer a lot more than that.* Obtenido de Linkedin.com: https://www.linkedin.com/pulse/api-management-platforms-comparative-study-awais-omer/

Pérez, J., & González, L. (2020). *Gobierno Electrónico y Automatización: El Caso de Estonia*. Aranzadi.

Pérez, J., & Hernández, F. (2023). *Automatización y Eficiencia en el Sector Jurídico: Herramientas y Tendencias*. Editorial Reus.

Pérez, J., & Hernández, F. (2023). *IA y Automatización en el Sector Legal: Desafíos y Oportunidades*. Tirant lo Blanch.

Pérez, J., & Hernández, F. (2023). *Innovación y Precisión en el Derecho: El Rol de la IA en la Práctica Jurídica Moderna*. Tirant lo Blanch.

Porter, M. E., & Heppelmann, J. E. (2021). *Smart, Connected Products: The Rise of the Data-Driven Product*. Harvard Business Review Press.

Ramirez Martinez, W. O. (2021). Asistente virtual para apoyar la atención al cliente en consultas legales de ámbito laboral en la empresa abogados Romero.

Redis. (3 de Octubre de 2023). *Introduction to Retrieval Augmented Generation (RAG)*. Obtenido de Redis: https://redis.io/glossary/retrieval-augmented-generation/

Robles, A., & Martínez, C. (2022). *Ética y Regulación de la Inteligencia Artificial en el Sector Público*. Editorial Reus.

Rodríguez, A., & Gómez, S. (2023). *Automatización Legal: El Rol del NLP en la Práctica Jurídica Moderna*. Aranzadi.

Rodríguez, L., & Pérez, C. (2021). *Inteligencia Artificial y Derecho: Herramientas para la Automatización Legal*. Aranzad.

Rodríguez, L., & Pérez, C. (2021). *Tecnologías de la Información en la Era Digital: El Papel de las APIs*. Aranzadi.

S/f. (30 de 08 de 2024). *Codigofacilito.com*. Obtenido de https://codigofacilito.com/articulos/mvc-model-view-controller-explicado

Sproat, R. (1992). *Morphology and Computation*. MIT Press.

Susskind, R. (2020). *Online Courts and the Future of Justice*. Oxford University Press.

Uscamayta, G. E. (2021). INTELIGENCIA ARTIFICIAL COMO HERRAMIENTA PARA MEJORAR LA CALIDAD Y CELERIDAD DE DECISIONES JURISDICCIONALES. *Tesis*. Universidad Andina del Cusco, Cusco.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., . . . & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems. En *Attention is All You Need. Advances in Neural Information Processing Systems* (págs. 30, 5998-6008.).

Zendesk. (19 de febrero de 2021). *¿Qué es la metodología ágil? ¿Para qué sirve?* Obtenido de https://www.zendesk.com.mx/blog/metodologia-agil-que-es/

Apéndice A

A.1. Repositorio (código fuente)

El código fuente y los datos utilizados en este trabajo están disponibles en el repositorio público de GitHub: https://github.com/AlexanderTS1/Tesis_Informes_Legales_LLM

Este repositorio contiene:

- El código correspondiente al pipeline, incluyendo la extracción de antecedentes y la generación de informe legal
- Los documentos legales utilizados en los experimentos
- Los resultados obtenidos en la generación de opiniones legales.

A.2. Evaluación Claude

Las evaluaciones realizadas por Claude sobre opiniones legales generados tanto por Llama, Mistral como Claude pueden encontrarse en el siguiente documento: https://docs.google.com/document/d/1TRgdKza6QfpMkzqjd38GxQbmfH0JLUiqegFRFRDjZMo/edit?usp=sharing

A.3. Ficha de evaluación

https://github.com/AlexanderTS1/Tesis_Informes_Legales_LLM/blob/main/FICHA%20DE%20EVALUACI%C3%93N.docx

A.4. Validación de ficha de evaluación

https://github.com/AlexanderTS1/Tesis_Informes_Legales_LLM/blob/main/VALIDACION%20FICHA%20DE%20EVALUACION.pdf

A.5 Fichas evaluadas por especialistas jurídicos

FICHA DE EVALUACIÓN

Proyecto de Investigación: **"AUTOMATIZACIÓN DE PRE-DICTÁMENES JURÍDICOS CON IA EN EL ÁREA DE ASESORÍA LEGAL DE LA MUNICIPALIDAD DE CALCA".**

Entidad: Municipalidad Provincial de Calca

Área: Asesor Legal - Especialista en Gerencia Municipal

I. DATOS DEL EVALUADOR

- Nombre del abogado evaluador: Jose Crisfian Montes Vargas
- Especialidad jurídica: Derecho Administrativo
- Años de experiencia profesional: 07 años
- Fecha de evaluación: 28 / 04 / 2026

II. INSTRUCCIONES

Lea el pre dictamen jurídico generado por el sistema y evalúe cada criterio asignando un puntaje de acuerdo con la siguiente escala. Considere la calidad del documento generado por el sistema antes de cualquier modificación o corrección posterior.

Escala de calificación:

Puntaje	Valoración
1	Muy deficiente
2	Deficiente
3	Regular
4	Bueno
5	Excelente

III. CRITERIOS DE EVALUACIÓN

Dimensión 1: Calidad Técnica y Coherencia Jurídica

(Metodología de Calidad de Software Público)

1.1. Coherencia jurídica del análisis

¿El razonamiento jurídico planteado en el pre dictamen es lógico, estructurado y consistente?

Puntaje (1–5): [5]

Jose Crisfian Montes Vargas
ABOGADO
C.A.A. 12933

1.2. Adecuación normativa

¿Las normas y jurisprudencia (v.g., SERVIR, TC, pronunciamientos) son pertinentes, vigentes y correctamente aplicadas?

Puntaje (1–5): [5]

1.3. Precisión en la interpretación de los antecedentes

¿El sistema procesa e interpreta correctamente los hechos, fechas de registro y actuados del expediente administrativo?

Puntaje (1–5): [4]

1.4. Fundamentación argumentativa

¿Las conclusiones y recomendaciones legales se derivan lógicamente de los considerandos y están debidamente sustentadas?

Puntaje (1–5): [4]

1.5. Estructura formal y redacción técnica

¿El documento respeta la estructura formal de un informe legal (Visto, Considerando, Opinión) y emplea un lenguaje claro y técnico?

Puntaje (1–5): [5]

Dimensión 2: Usabilidad e Interacción con el Usuario

(Metodología de Heurísticas de Nielsen)

2.1. Facilidad de operación e interfaz

¿El sistema es intuitivo para el operador legal y facilita la carga de archivos o parámetros de búsqueda sin requerir asistencia técnica constante?

Puntaje (1–5): [5]

2.2. Nivel de autonomía del sistema

¿El borrador del informe legal generado requiere un nivel mínimo de modificaciones o correcciones manuales antes de su visación?

Puntaje (1–5): [4]

José Cristian Montes Vargas
ABOGADO
C.A.A. 12633

Dimensión 3: Impacto Organizacional

(Metodología de Impacto y Productividad)

3.1. Optimización del tiempo de respuesta

¿Considera que el uso del sistema reduce significativamente las horas-hombre dedicadas a la redacción estándar de un pre dictamen?

Puntaje (1–5): [5]

3.2. Utilidad como herramienta de apoyo

¿Considera que el sistema es viable y útil para ser institucionalizado como herramienta de soporte en la Oficina de Asesoría Jurídica?

Puntaje (1–5): [5]

IV. EVALUACIÓN GLOBAL

En términos generales, ¿cómo califica el desempeño e impacto del sistema en la gestión jurídica local?

Puntaje (1–5): [5]

V. OBSERVACIONES CUALITATIVAS

Observaciones y recomendaciones del evaluador:

Tener Cuidado con fundamentación argumentativa

Jose Cristian Montes Vargas
ABOGADO
C.A.A. 12633

FICHA DE EVALUACIÓN

Proyecto de Investigación: **"AUTOMATIZACIÓN DE PRE-DICTÁMENES JURÍDICOS CON IA EN EL ÁREA DE ASESORÍA LEGAL DE LA MUNICIPALIDAD DE CALCA".**

Entidad: Municipalidad Provincial de Calca

Área: Especialista PAD

I. DATOS DEL EVALUADOR

- Nombre del abogado evaluador: Jhoel Mormontoy Torres
- Especialidad jurídica: Derecho Administrativo
- Años de experiencia profesional: 04 años
- Fecha de evaluación: 28 / 04 / 2026

II. INSTRUCCIONES

Lea el pre dictamen jurídico generado por el sistema y evalúe cada criterio asignando un puntaje de acuerdo con la siguiente escala. Considere la calidad del documento generado por el sistema antes de cualquier modificación o corrección posterior.

Escala de calificación:

Puntaje	Valoración
1	Muy deficiente
2	Deficiente
3	Regular
4	Bueno
5	Excelente

III. CRITERIOS DE EVALUACIÓN

Dimensión 1: Calidad Técnica y Coherencia Jurídica

(Metodología de Calidad de Software Público)

1.1. Coherencia jurídica del análisis

¿El razonamiento jurídico planteado en el pre dictamen es lógico, estructurado y consistente?

Puntaje (1–5): [5]

Jhoel Mormontoy Torres
ABOGADO
CAL-SUR N° 02289

1.2. Adecuación normativa

¿Las normas y jurisprudencia (v.g., SERVIR, TC, ¿pronunciamientos) son pertinentes, vigentes y correctamente aplicadas?

Puntaje (1–5): [4]

1.3. Precisión en la interpretación de los antecedentes

¿El sistema procesa e interpreta correctamente los hechos, fechas de registro y actuados del expediente administrativo?

Puntaje (1–5): [4]

1.4. Fundamentación argumentativa

¿Las conclusiones y recomendaciones legales se derivan lógicamente de los considerandos y están debidamente sustentadas?

Puntaje (1–5): [4]

1.5. Estructura formal y redacción técnica

¿El documento respeta la estructura formal de un informe legal (Visto, Considerando, Opinión) y emplea un lenguaje claro y técnico?

Puntaje (1–5): [5]

Dimensión 2: Usabilidad e Interacción con el Usuario

(Metodología de Heurísticas de Nielsen)

2.1. Facilidad de operación e interfaz

¿El sistema es intuitivo para el operador legal y facilita la carga de archivos o parámetros de búsqueda sin requerir asistencia técnica constante?

Puntaje (1–5): [5]

2.2. Nivel de autonomía del sistema

¿El borrador del informe legal generado requiere un nivel mínimo de modificaciones o correcciones manuales antes de su visación?

Puntaje (1–5): [4]

Jhoel Mormontoy Torres
ABOGADO
CAL-SUR N° 02289

Dimensión 3: Impacto Organizacional

(Metodología de Impacto y Productividad)

3.1. Optimización del tiempo de respuesta

¿Considera que el uso del sistema reduce significativamente las horas-hombre dedicadas a la redacción estándar de un pre dictamen?

Puntaje (1–5): [5]

3.2. Utilidad como herramienta de apoyo

¿Considera que el sistema es viable y útil para ser institucionalizado como herramienta de soporte en la Oficina de Asesoría Jurídica?

Puntaje (1–5): [5]

IV. EVALUACIÓN GLOBAL

En términos generales, ¿cómo califica el desempeño e impacto del sistema en la gestión jurídica local?

Puntaje (1–5): [4]

V. OBSERVACIONES CUALITATIVAS

Observaciones y recomendaciones del evaluador:

Ninguna

Jhoel Mormontoy Torres
ABOGADO
CAL-SUR N° 02289

FICHA DE EVALUACIÓN

Proyecto de Investigación: **"AUTOMATIZACIÓN DE PRE-DICTÁMENES JURÍDICOS CON IA EN EL ÁREA DE ASESORÍA LEGAL DE LA MUNICIPALIDAD DE CALCA".**

Entidad: MUNICIPALIDAD PROVINCIAL DE CALCA

Área: OFICINA DE ASESORIA JURIDICA

I. DATOS DEL EVALUADOR

- Nombre del abogado evaluador: JOEL ORNACHEA LARA
- Especialidad jurídica: DERECHO ADMINISTRATIVO Y GESTION PUBLICA
- Años de experiencia profesional: 4 AÑOS
- Fecha de evaluación: 28 / 04 / 26

II. INSTRUCCIONES

Lea el pre dictamen jurídico generado por el sistema y evalúe cada criterio asignando un puntaje de acuerdo con la siguiente escala. Considere la calidad del documento generado por el sistema antes de cualquier modificación o corrección posterior.

Escala de calificación:

Puntaje	Valoración
1	Muy deficiente
2	Deficiente
3	Regular
4	Bueno
5	Excelente

III. CRITERIOS DE EVALUACIÓN

Dimensión 1: Calidad Técnica y Coherencia Jurídica

(Metodología de Calidad de Software Público)

1.1. Coherencia jurídica del análisis

¿El razonamiento jurídico planteado en el pre dictamen es lógico, estructurado y consistente?

Puntaje (1–5): [4]

MUNICIPALIDAD PROVINCIAL DE CALCA
OFICINA DE ASESORÍA JURÍDICA
Abog. Joel Ornachea Lara
ASESOR JURÍDICO

1.2. Adecuación normativa

¿Las normas y jurisprudencia (v.g., SERVIR, TC, pronunciamientos) son pertinentes, vigentes y correctamente aplicadas?

Puntaje (1–5): [4]

1.3. Precisión en la interpretación de los antecedentes

¿El sistema procesa e interpreta correctamente los hechos, fechas de registro y actuados del expediente administrativo?

Puntaje (1–5): [4]

1.4. Fundamentación argumentativa

¿Las conclusiones y recomendaciones legales se derivan lógicamente de los considerandos y están debidamente sustentadas?

Puntaje (1–5): [4]

1.5. Estructura formal y redacción técnica

¿El documento respeta la estructura formal de un informe legal (Visto, Considerando, Opinión) y emplea un lenguaje claro y técnico?

Puntaje (1–5): [4]

Dimensión 2: Usabilidad e Interacción con el Usuario

(Metodología de Heurísticas de Nielsen)

2.1. Facilidad de operación e interfaz

¿El sistema es intuitivo para el operador legal y facilita la carga de archivos o parámetros de búsqueda sin requerir asistencia técnica constante?

Puntaje (1–5): [5]

2.2. Nivel de autonomía del sistema

¿El borrador del informe legal generado requiere un nivel mínimo de modificaciones o correcciones manuales antes de su visación?

Puntaje (1–5): [4]

MUNICIPALIDAD PROVINCIAL DE CALCA
OFICINA DE ASESORÍA JURÍDICA
Abog. Joel Ormachea Lara
ASESOR JURÍDICO

Dimensión 3: Impacto Organizacional

(Metodología de Impacto y Productividad)

3.1. Optimización del tiempo de respuesta

¿Considera que el uso del sistema reduce significativamente las horas-hombre dedicadas a la redacción estándar de un pre dictamen?

Puntaje (1–5): [5]

3.2. Utilidad como herramienta de apoyo

¿Considera que el sistema es viable y útil para ser institucionalizado como herramienta de soporte en la Oficina de Asesoría Jurídica?

Puntaje (1–5): [5]

IV. EVALUACIÓN GLOBAL

En términos generales, ¿cómo califica el desempeño e impacto del sistema en la gestión jurídica local?

Puntaje (1–5): [4]

V. OBSERVACIONES CUALITATIVAS

Observaciones y recomendaciones del evaluador:

SE DEBE CUIDAR COHERENCIA JURIDICA DEL ANALISIS DEL CASO

MUNICIPALIDAD PROVINCIAL DE CALCA
OFICINA DE ASESORÍA JURÍDICA
Abog. Joel Ormachea Lara
ASESOR JURÍDICO

FICHA DE EVALUACIÓN

Proyecto de Investigación: **"AUTOMATIZACIÓN DE PRE-DICTÁMENES JURÍDICOS CON IA EN EL ÁREA DE ASESORÍA LEGAL DE LA MUNICIPALIDAD DE CALCA".**

Entidad: Municipalidad Provincial de Calca

Área: Oficina de Asesoría Legal

I. DATOS DEL EVALUADOR

- Nombre del abogado evaluador: Betsabe Arriola Arriola
- Especialidad jurídica: Derecho Administrativo y Gestión pública
- Años de experiencia profesional: 05 años
- Fecha de evaluación: 28 / 04 / 2026

II. INSTRUCCIONES

Lea el pre dictamen jurídico generado por el sistema y evalúe cada criterio asignando un puntaje de acuerdo con la siguiente escala. Considere la calidad del documento generado por el sistema antes de cualquier modificación o corrección posterior.

Escala de calificación:

Puntaje	Valoración
1	Muy deficiente
2	Deficiente
3	Regular
4	Bueno
5	Excelente

III. CRITERIOS DE EVALUACIÓN

Dimensión 1: Calidad Técnica y Coherencia Jurídica

(Metodología de Calidad de Software Público)

1.1. Coherencia jurídica del análisis

¿El razonamiento jurídico planteado en el pre dictamen es lógico, estructurado y consistente?

Puntaje (1–5): [5]

Betsabe Arriola Arriola
ABOGADO
C.A.C. 7617

1.2. Adecuación normativa

¿Las normas y jurisprudencia (v.g., SERVIR, TC, pronunciamientos) son pertinentes, vigentes y correctamente aplicadas?

Puntaje (1–5): [5]

1.3. Precisión en la interpretación de los antecedentes

¿El sistema procesa e interpreta correctamente los hechos, fechas de registro y actuados del expediente administrativo?

Puntaje (1–5): [5]

1.4. Fundamentación argumentativa

¿Las conclusiones y recomendaciones legales se derivan lógicamente de los considerandos y están debidamente sustentadas?

Puntaje (1–5): [5]

1.5. Estructura formal y redacción técnica

¿El documento respeta la estructura formal de un informe legal (Visto, Considerando, Opinión) y emplea un lenguaje claro y técnico?

Puntaje (1–5): [5]

Dimensión 2: Usabilidad e Interacción con el Usuario

(Metodología de Heurísticas de Nielsen)

2.1. Facilidad de operación e interfaz

¿El sistema es intuitivo para el operador legal y facilita la carga de archivos o parámetros de búsqueda sin requerir asistencia técnica constante?

Puntaje (1–5): [5]

2.2. Nivel de autonomía del sistema

¿El borrador del informe legal generado requiere un nivel mínimo de modificaciones o correcciones manuales antes de su visación?

Puntaje (1–5): [5]

Betsabe Arnola Arnola
ABOGADO
C.A.C 7617

Dimensión 3: Impacto Organizacional

(Metodología de Impacto y Productividad)

3.1. Optimización del tiempo de respuesta

¿Considera que el uso del sistema reduce significativamente las horas-hombre dedicadas a la redacción estándar de un pre dictamen?

Puntaje (1–5): [5]

3.2. Utilidad como herramienta de apoyo

¿Considera que el sistema es viable y útil para ser institucionalizado como herramienta de soporte en la Oficina de Asesoría Jurídica?

Puntaje (1–5): [5]

IV. EVALUACIÓN GLOBAL

En términos generales, ¿cómo califica el desempeño e impacto del sistema en la gestión jurídica local?

Puntaje (1–5): [5]

V. OBSERVACIONES CUALITATIVAS

Observaciones y recomendaciones del evaluador:

Betsabe Arriola Arriola
ABOGADO
C.A.C 7617

FICHA DE EVALUACIÓN

Proyecto de Investigación: **"AUTOMATIZACIÓN DE PRE-DICTÁMENES JURÍDICOS CON IA EN EL ÁREA DE ASESORÍA LEGAL DE LA MUNICIPALIDAD DE CALCA".**

Entidad: Municipalidad Provincial de Calca

Área: Ejecucion Coactiva.

I. DATOS DEL EVALUADOR

- Nombre del abogado evaluador: Estefany Aucca Enriquez.
- Especialidad jurídica: Derecho Administrativo y Gestion Publica
- Años de experiencia profesional: 05 años
- Fecha de evaluación: 29 / 04 / 2026

II. INSTRUCCIONES

Lea el pre dictamen jurídico generado por el sistema y evalúe cada criterio asignando un puntaje de acuerdo con la siguiente escala. Considere la calidad del documento generado por el sistema antes de cualquier modificación o corrección posterior.

Escala de calificación:

Puntaje	Valoración
1	Muy deficiente
2	Deficiente
3	Regular
4	Bueno
5	Excelente

III. CRITERIOS DE EVALUACIÓN

Dimensión 1: Calidad Técnica y Coherencia Jurídica

(Metodología de Calidad de Software Público)

1.1. Coherencia jurídica del análisis

¿El razonamiento jurídico planteado en el pre dictamen es lógico, estructurado y consistente?

Puntaje (1–5): [5]

Abg. Estefany J. Aucca Enriquez
ICAC. 9784

1.2. Adecuación normativa

¿Las normas y jurisprudencia (v.g., SERVIR, TC, pronunciamientos) son pertinentes, vigentes y correctamente aplicadas?

Puntaje (1–5): [5]

1.3. Precisión en la interpretación de los antecedentes

¿El sistema procesa e interpreta correctamente los hechos, fechas de registro y actuados del expediente administrativo?

Puntaje (1–5): [4]

1.4. Fundamentación argumentativa

¿Las conclusiones y recomendaciones legales se derivan lógicamente de los considerandos y están debidamente sustentadas?

Puntaje (1–5): [4]

1.5. Estructura formal y redacción técnica

¿El documento respeta la estructura formal de un informe legal (Visto, Considerando, Opinión) y emplea un lenguaje claro y técnico?

Puntaje (1–5): [5]

Dimensión 2: Usabilidad e Interacción con el Usuario

(Metodología de Heurísticas de Nielsen)

2.1. Facilidad de operación e interfaz

¿El sistema es intuitivo para el operador legal y facilita la carga de archivos o parámetros de búsqueda sin requerir asistencia técnica constante?

Puntaje (1–5): [5]

2.2. Nivel de autonomía del sistema

¿El borrador del informe legal generado requiere un nivel mínimo de modificaciones o correcciones manuales antes de su visación?

Puntaje (1–5): [4]

Abg. Estefany J. Aucca Enriquez
ICAC. 9784

Dimensión 3: Impacto Organizacional

(Metodología de Impacto y Productividad)

3.1. Optimización del tiempo de respuesta

¿Considera que el uso del sistema reduce significativamente las horas-hombre dedicadas a la redacción estándar de un pre dictamen?

Puntaje (1–5): [5]

3.2. Utilidad como herramienta de apoyo

¿Considera que el sistema es viable y útil para ser institucionalizado como herramienta de soporte en la Oficina de Asesoría Jurídica?

Puntaje (1–5): [5]

IV. EVALUACIÓN GLOBAL

En términos generales, ¿cómo califica el desempeño e impacto del sistema en la gestión jurídica local?

Puntaje (1–5): [5]

V. OBSERVACIONES CUALITATIVAS

Observaciones y recomendaciones del evaluador:

NINGUNA.

Abg. Estefany J. Aucca Enriquez
ICAC. 9784

FICHA DE EVALUACIÓN

Proyecto de Investigación: **"AUTOMATIZACIÓN DE PRE-DICTÁMENES JURÍDICOS CON IA EN EL ÁREA DE ASESORÍA LEGAL DE LA MUNICIPALIDAD DE CALCA".**

Entidad: Municipalidad Provincial de Calca

Área: Procuraduría Pública Municipal

I. DATOS DEL EVALUADOR

- Nombre del abogado evaluador: Frank Yuri Negron Torres
- Especialidad jurídica: Derecho Administrativo
- Años de experiencia profesional: 07 años
- Fecha de evaluación: 29 / 04 / 2026

FRANK YURI NEGRON TORRES
ABOGADO
Reg 8200

II. INSTRUCCIONES

Lea el pre dictamen jurídico generado por el sistema y evalúe cada criterio asignando un puntaje de acuerdo con la siguiente escala. Considere la calidad del documento generado por el sistema antes de cualquier modificación o corrección posterior.

Escala de calificación:

Puntaje	Valoración
1	Muy deficiente
2	Deficiente
3	Regular
4	Bueno
5	Excelente

III. CRITERIOS DE EVALUACIÓN

Dimensión 1: Calidad Técnica y Coherencia Jurídica

(Metodología de Calidad de Software Público)

1.1. Coherencia jurídica del análisis

¿El razonamiento jurídico planteado en el pre dictamen es lógico, estructurado y consistente?

Puntaje (1–5): [4]

1.2. Adecuación normativa

¿Las normas y jurisprudencia (v.g., SERVIR, TC, pronunciamientos) son pertinentes, vigentes y correctamente aplicadas?

Puntaje (1–5): [4]

1.3. Precisión en la interpretación de los antecedentes

¿El sistema procesa e interpreta correctamente los hechos, fechas de registro y actuados del expediente administrativo?

Puntaje (1–5): [4]

1.4. Fundamentación argumentativa

¿Las conclusiones y recomendaciones legales se derivan lógicamente de los considerandos y están debidamente sustentadas?

Puntaje (1–5): [4]

1.5. Estructura formal y redacción técnica

¿El documento respeta la estructura formal de un informe legal (Visto, Considerando, Opinión) y emplea un lenguaje claro y técnico?

Puntaje (1–5): [5]

FRANK YURI NEGRON TORRES
ABOGADO
Reg. 8200

Dimensión 2: Usabilidad e Interacción con el Usuario

(Metodología de Heurísticas de Nielsen)

2.1. Facilidad de operación e interfaz

¿El sistema es intuitivo para el operador legal y facilita la carga de archivos o parámetros de búsqueda sin requerir asistencia técnica constante?

Puntaje (1–5): [5]

2.2. Nivel de autonomía del sistema

¿El borrador del informe legal generado requiere un nivel mínimo de modificaciones o correcciones manuales antes de su visación?

Puntaje (1–5): [4]

Dimensión 3: Impacto Organizacional

(Metodología de Impacto y Productividad)

3.1. Optimización del tiempo de respuesta

¿Considera que el uso del sistema reduce significativamente las horas-hombre dedicadas a la redacción estándar de un pre dictamen?

Puntaje (1–5): [5]

3.2. Utilidad como herramienta de apoyo

¿Considera que el sistema es viable y útil para ser institucionalizado como herramienta de soporte en la Oficina de Asesoría Jurídica?

Puntaje (1–5): [5]

IV. EVALUACIÓN GLOBAL

En términos generales, ¿cómo califica el desempeño e impacto del sistema en la gestión jurídica local?

Puntaje (1–5): [4]

V. OBSERVACIONES CUALITATIVAS

Observaciones y recomendaciones del evaluador:

FRANK YURI NEGRON TORRES
ABOGADO
Reg 8200