



**UIE - GISI**

# CertyLex

AI Legal Assistant

Manuel Mateo Delgado-Gambino López

# Index

**01** Introduction

**02** Abstract + Problem Statement

**03** Objectives + Scope

**04** Pipeline

**05** App + Models + Architecture

**06** Data Preprocessing

**07** Vision

**08** Conclusion

# Introduction



## **Abstract**

Manual contract review is slow and error-prone

Scanned documents need OCR before analysis

Legal teams spend hours on summaries, editing and risk checking

# Solution Overview

## Multimodal AI Pipeline

- Clean text
- Summarize content
- Flag risks
- Improve phrasing
- Image OCR management

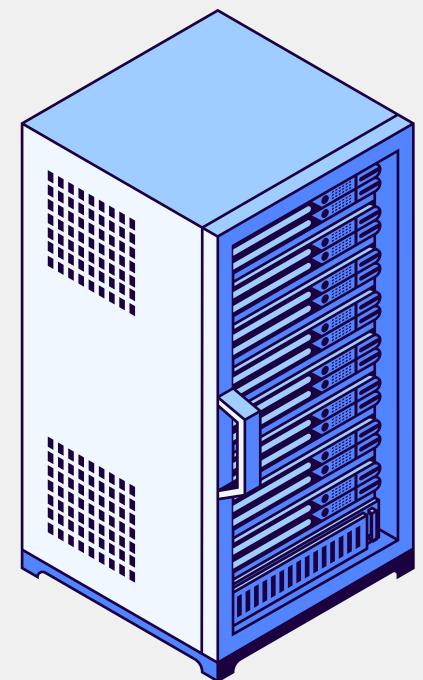
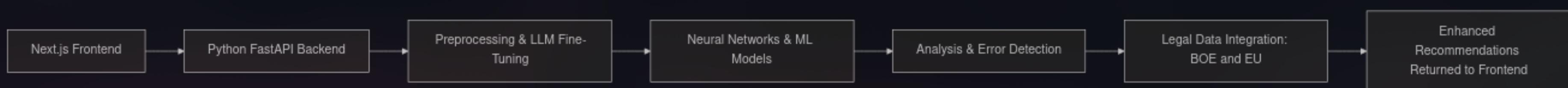
## Core Benefits

- Speed of procedure
- Improved accuracy
- Check regulatory oversights

## User Workflow

Upload PDF/image →  
AI analyzes & suggests →  
Export formalized,  
client-ready contract.

# Project Pipeline



# Flowchart

Text styling and preprocessing

Regex and data processing

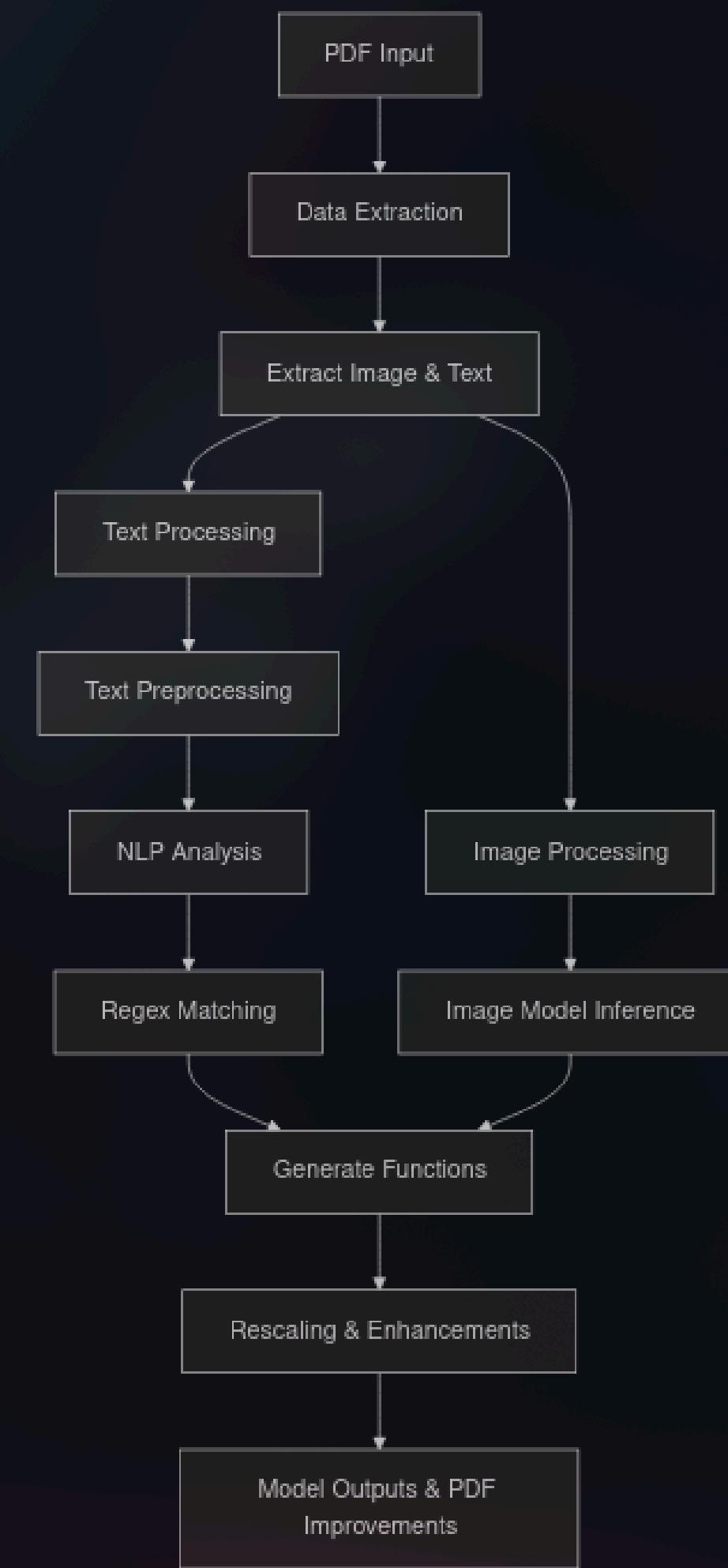
Text summarization

Responses and enhancements

Image data extraction

Image processing

Image output



# Models

**DeepSeek-R1:32b**

**DeepSeek-R1:8b - custom**

**Llama3.2:3b**

**LLama3.2-vision:11b**

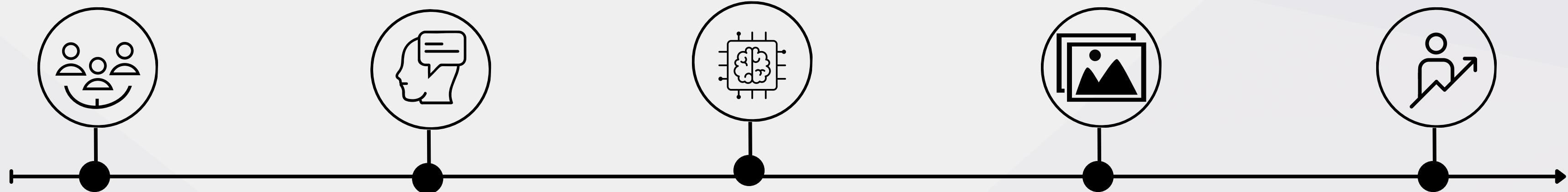
**BOE\_BART**

**BOE\_GPT2**



# App

# App Overview



## User

The user land on the agent section of the app and he can ask questions to the AI Agent.

## AI Agent

The agent is trained to respond only in spanish and to questions related to the legal field.

## PDF Analyzer

The agent is capable to extract text information about a pdf file and give recommendations and search for frauds.

## Image Upscaler

This app also is capable of detect any image inside the PDF and returns data about the image and generates improvements on the image.

## Future Implementations

Right now there is no user management and it also doesnt perform the images as expected

# Chat Section

The screenshot shows the CertyLex AI Legal Assistant interface running in a web browser (Mozilla Firefox) on a Linux desktop. The browser window title is "CertyLex". The interface includes a top navigation bar with tabs for "CertyLex" (selected), "AI Legal Assistant", "Chat General" (selected), "Análisis de Documentos", "Mejora de Imágenes", and "Mejora con CNN". A "Analizar PDF" button is also present. On the left, there's a sidebar titled "Configuración" with sections for "Información" (describing legal information from official sources like the BOE), "Model Parameters" (Temperature: 0.7, Creativity slider, Precise/Creative, Top-p: 1, Diversity slider, Max tokens: 1024, Response length slider), and "Fuentes Consultadas" (listing the Boletín Oficial del Estado, Código Civil Español, Ley General Tributaria, and Leyes Orgánicas del Estado). The main content area displays the "Asistente Legal CertyLex" component, which is described as an AI legal assistant with access to official Spanish state information. It shows a dropdown for "Modelo: deepseek-r1:32b". Below it is a large input field containing placeholder text: "Consulta información legal basada en el Boletín Oficial del Estado". A "Sugerencias" section at the bottom provides three prompts: "Generame un contrato formal entre 2 partes.", "Generame un reporte de ventas generadas por varios vendedores.", and "Generame una baja de un empleado por enfermedad.". The bottom right of the input field has a small "Up" arrow icon. The footer of the browser window displays system status icons and the text "© 2025 CertyLex. All rights reserved."

## Selección de Modelo

- deepseek-r1:32b (19851337640)
- deepseek-r1:8b (4920738407)
- llama3.2:3b (2019393189)
- llama3.2-vision:11b (7901829417)

Cancelar

Seleccionar

A. The model 32b was only used at testing due to low ram capacity (30.5GB) on my machine.

B. The model Llama Vision is not the best one for text treatment due to ram consumption.

C. The base deepseekr1:8b tends to hallucinate and returns responses in various languages

D. The llama3.2:3b responds on a better/ formal spanish but the responses are basic and short.

Modelo: deepseek-r1:8b ▾

Contrato Formal entre un Empleador de Empresa Técnica y un Desarrollador Junior

Partes Interesadas:

- Employer: Tech Innovators Inc., ubicada en Silicon Valley, California.
- Employee: Juan Garcia, un desarrollador junior basado en San Francisco, California.

Duración del Empleo:

Este contrato surt efecto por un año, comenzando el [Fecha de Inicio], y renovable por acuerdo mutuo.

Compensación:

- Salario base: \$60,000 anualmente, pagado mensualmente a través de cheques semimensuales.
- Bonificaciones: 10% del salario base upon completion of projects.
- Beneficios estándar de la empresa, como seguros de salud y planes de retiro.

Responsabilidades del Trabajo:

- Desarrollar software según sea asignado por el jefe de proyectos.
- Utilizar el juicio mejor en cumplimiento con las políticas de la empresa.

hola me puedes generar un contrato formal entre un empleador de una empresa tecnologica y un junior developer?

# BOE BART && BOE GPT2

BOE\_GPT2

## Purpose:

- Generative model based on GPT-2 fine-tuned on BOE documents
  - Generates new legal texts with official document style and format
  - Continues text from prompts (e.g., resolution titles or legal headers)

## **Limitations:**

- High computational resources required for training
  - Slow generation of extensive texts
  - Can produce legally incorrect or fictional information
  - Limited context window (512 tokens)
  - Coherence issues in longer documents
  - Quality depends on training dataset representativeness

BOE BART

## Purpose:

- Model designed for summarizing BOE documents
  - Uses BART architecture for conditional generation or abstractive summarization
  - Complements genetic algorithm-based extractive summarization

### **Limitations:**

- Unsupervised training restricts accurate summarization capabilities
  - Lacks labeled summary corpus for supervised learning
  - More complex training than extractive approaches
  - May generate information not present in original document
  - Limited by context window size (1024 tokens)
  - Difficult to evaluate legal accuracy of generated summaries

### › GPT-2 Output:

BART Output:

Describe en detalle las principales características de la Constitución Española, incluyendo su estructura, principios fundamentales y derechos reconocidos.

# Output Deepseek-r1:8b - custom

A ~ Contrato Formal entre Empleador de Empresa Técnica y Desarrollador Junior

Partes Interesadas:

Empleador: Tech Innovators Inc., ubicada en Silicon Valley, California.  
Empleado: Juan Garcia, un desarrollador junior basado en San Francisco, California.

Duración del Empleo:  
Este contrato surt efecto por un año, comenzando el [Fecha de Inicio], y renovable de manera mutual after agreed upon.

Compensación:

Salario base: \$60,000 anualmente, pagado mensualmente a través de cheques semiprofesionales.  
Bonificaciones: 10% del salario baseupon completar un proyecto.  
Beneficios estándar como seguro de salud, planes de jubilación y más.

Responsabilidades del Trabajo:

Desarrollar software según asignado por el gerente.  
Usar el juicio propio en cumplimiento con las políticas de la empresa.  
Mantener la confidencialidad de la información de la compañía.

Confidencialidad:  
Ambas partes acuerdan no revelar secretos de la empresa ni compartir información externamente por un período de dos años después del despido.

Propiedad Intelectual:  
El empleado tiene derecho a las inversiones creadas fuera de las horas laborales a menos que se especifique lo contrario por escrito.

Terminación:  
Ambas partes pueden rescindir el contrato con 90 días de notificación escrita. La compañía puede otorgar un pago de despido de tres meses de salario.

Conflictos de Intereses:  
El empleado debe informar al empleador si está involucrado en una empresa competidora o tomando clientes.

Ley Aplicable:  
La ley de California se aplica a todos los términos y condiciones.

Acuerdo Completo:  
No existen antecedentes previos que tengan prevalencia sobre este contrato. Las modificaciones deben estar hechas por escrito con aprobación de ambas partes.

Resolución de Disputas:  
Primero, mediación, luego arbitraje y finalmente tribunal en caso necesario.

Notarización:  
Ambas partes deben firmar ante un notario público para validar el contrato.

Asesoramiento Legal:  
El empleador y el empleado deben consultar con abogados para asegurar que todo esté legal y equitativo.

Cláusulas Adicionales:

Alcance del Trabajo: Deliverables y cronogramas especificados, junto con métricas de evaluación trimestrales.  
Proceso de Evaluación: Reuniones cuatrimestrales y establecimiento de metas por parte conjunta.  
Proceso de Modificaciones: Aprobación escrita de ambas partes requerida para cambios.  
Acciones Disciplinarias: Periodo de tres días para responder a violaciones del empleado.

Aceptación:  
Ambas partes acuerdan con los términos delineados en este contrato al firmar la misma.

This custom model solves some issues that default deepseek has:

1. All text is parsed to spanish
2. Improves styling and format
3. Less likely to hallucinations

Customozed via:

1. Modelfile
2. Prompting
3. Fine-tuned with BOE dataset

# PDF Analysis Section

The screenshot shows a web browser window for the CertyLex AI Legal Assistant. The URL is <http://localhost:3000/pdf-analysis>. The page has a dark theme with purple and white highlights. At the top, there are tabs for "CertyLex AI Legal Assistant", "Chat General", "Análisis de Documentos", "Mejora de Imágenes", and "Mejora con CNN". A prominent blue button labeled "Analizar PDF" is located on the right. On the left, there's a sidebar with sections for "Configuración", "Análisis de Documentos", and "Fuentes Consultadas". The main content area is titled "Análisis de Documentos Legales" and contains a dashed box for uploading a PDF file, with a "Seleccionar PDF" button below it.

1 3 4 5 9 A CertyLex

7% 9.57GB 62°C 37°C 38°C 03:26 AM 0.0B/s ± 0.0B/s 25 15 12°C | Santiago De Compostela, Spain

Presentación Diapositivas Prx CertyLex

http://localhost:3000/pdf-analysis

110%

CertyLex AI Legal Assistant Chat General Análisis de Documentos Mejora de Imágenes Mejora con CNN Analizar PDF

Configuración

Análisis de Documentos Legales

Sube un documento PDF para obtener análisis y mejoras

Sube un documento para analizar

Seleccionar PDF

Análisis de Documentos

Sube un documento PDF para que CertyLex lo analice, identifique mejoras y proporcione recomendaciones legales basadas en fuentes oficiales.

Fuentes Consultadas

- Boletín Oficial del Estado
- Código Civil Español
- Ley General Tributaria
- Leyes Orgánicas del Estado

© 2025 CertyLex. All rights reserved.

# Models

MODEL	RAM	QUALITIES	DOWNSIDES
Deepseek-r1:32b	14.6GiB	Best Performance	Slow Performance
Deepseek-r1:8b	3.7GiB	Good Performance	Speaks Chinese
Llama3.2:3b	2.1GiB	Fast Performance	Narrow Context Window
LLama3.2-vision:11b	8.3GiB	Best Suited for Images	Slow for Text Processing
Deepseek-r1:8b - custom	3.9GiB	Best Suited for Analysis	Modify User Prompt

# PDF Analysis Section

The screenshot shows the CertyLex AI Legal Assistant interface, specifically the PDF Analysis section. The top navigation bar includes tabs for Chat General, Análisis de Documentos, Mejora de Imágenes, Mejora con CNN, and a prominent blue button labeled "Analizar PDF". On the left, there's a sidebar with sections for "Configuración" (Analysis of Documents, Fuentes Consultadas), "Análisis de Documentos Legales" (Upload a PDF document for analysis and improvements), and a preview of the analyzed document content. The main content area displays the analyzed text of a professional services contract, highlighting entities like "DON JOSÉ GARCÍA LÓPEZ" and "DOÑA MARÍA RODRÍGUEZ SÁNCHEZ". To the right, there's a "Sugerencias" (Suggestions) section with three cards: "¿Qué mejoras sugerires para este contrato?", "Identifica posibles cláusulas abusivas", and "Resume los puntos principales del documento". A "Consulta sobre legislación española..." (Spanish legislation consultation) button is also present.

1 3 4 5 9 CertyLex

Presentación Diapositivas Pr CertyLex

http://localhost:3000/pdf-analysis

CertyLex AI Legal Assistant Chat General Análisis de Documentos Mejora de Imágenes Mejora con CNN Analizar PDF

Configuración

Análisis de Documentos Legales

Sube un documento PDF para obtener análisis y mejoras

Original Con Mejoras Análisis Entidades

CONTRATO DE PRESTACIÓN DE SERVICIOS PROFESIONALES

En Madrid, a 4 de Mayo de 2025

REUNIDOS

De una parte, DON JOSÉ GARCÍA LÓPEZ, mayor de edad, con domicilio en Calle Mayor 25, 28001 Madrid, y con DNI número 12345678A, actuando en su propio nombre y derecho (en adelante, el "CLIENTE").

Y de otra parte, DOÑA MARÍA RODRÍGUEZ SÁNCHEZ, mayor de edad, con domicilio profesional en Avenida de la Constitución 15, 28002 Madrid, y con DNI número 87654321B, actuando en su propio nombre y derecho (en adelante, el "PROFESIONAL").

Ambas partes (en adelante, conjuntamente, las "Partes"), reconociéndose mutuamente capacidad legal suficiente para contratar y obligarse en la representación que actúan y siendo responsables de la veracidad de sus manifestaciones,

EXPONEN

I. Que el CLIENTE está interesado en contratar los servicios profesionales del PROFESIONAL, consistentes en asesoramiento legal en materia contractual.

II. Que el PROFESIONAL tiene los conocimientos y la experiencia necesarios para prestar dichos servicios.

III. Que las Partes están interesadas en celebrar un contrato de prestación de servicios en virtud del cual el PROFESIONAL preste al CLIENTE los servicios que aquí se definen.

IV. Que las Partes han acordado otorgar el presente contrato de prestación de servicios profesionales (en adelante, el "Contrato"), con sujeción a las siguientes:

cm5ii8n1r0001nq0xq... Original

Consulta sobre el documento

Asistente Legal CertyLex

Consulta información legal basada en el Boletín Oficial del Estado

Sugerencias

¿Qué mejoras sugerires para este contrato?

Identifica posibles cláusulas abusivas

Resume los puntos principales del documento

Consulta sobre legislación española...

© 2025 CertyLex. All rights reserved.

# Section Overview

**Suggestions**

**Consulting**

**Analysis**

# Suggestions

cm5ii8n1r0001nq0xq... Análisis

Puntos clave identificados

**Sugerencia**

1. **\*\*Incluir una cláusula de cancelación\*\*:** Es importante incluir una cláusula que establezca las condiciones bajo las cuales el contrato puede ser resuelto por cualquiera de las partes, incluyendo la notificación previa y cualquier compensación que deba pagarse.

2. **\*\*Definir la naturaleza del servicio\*\*:** Aunque se menciona que los servicios son de asesoramiento legal en materia contractual, es importante definir con más detalle qué tipo de servicios se entienden por este término para evitar malentendidos.

3. **\*\*Establecer límites de confidencialidad\*\*:** La cláusula de secreto sobre la información recibida del CLIENTE puede ser demasiado amplia y no especificar qué tipos de información son confidenciales.

4. **\*\*Incluir una cláusula de disputas\*\*:** En lugar de someterse al arbitraje, es posible incluir una cláusula que establezca un Es importante definir qué condiciones deben cumplirse para continuar o suspender el contrato.

**Advertencia**

Esta cláusula podría considerarse abusiva al no permitir la resolución del contrato por ninguna de las partes.  
Se recomienda modificarla para incluir causas justificadas de resolución y preaviso razonable.

**Nota Legal**

Se recomienda especificar la corte de arbitraje y el procedimiento a seguir.

**A. Detects possible frauds or inconsistencies**

**B. Alerts the user about abusive or unfair clauses**

**C. Generates workarounds or fixes that could be applied to the document**

**D. Searches from all the pretrained data with metaheuristic algorithms to retrieve the most accurate response**

# Analysis and Improvements

CertyLex AI Legal Assistant Chat General Análisis de Documentos Mejora de Imágenes Mejora con CNN Analizar PDF

**Configuración**

Análisis de Documentos Legales  
Sube un documento PDF para obtener análisis y mejoras

Original Con Mejoras Análisis Entidades

Documento con Mejoras Sugeridas  
A continuación se muestra el documento con anotaciones y sugerencias de mejora. Las secciones con /\* SUGERENCIAS \*/ contienen recomendaciones específicas.

CONTRATO DE PRESTACIÓN DE SERVICIOS PROFESIONALES  
En Madrid, a 4 de Mayo de 2025  
REUNIDOS  
De una parte, DON JOSÉ GARCÍA LÓPEZ, mayor de edad, con domicilio en Calle Mayor 25, 28001 Madrid, y con DNI número 12345678A, actuando en su propio nombre y derecho (en adelante, el "CLIENTE").  
Y de otra parte, DOÑA MARÍA RODRÍGUEZ SÁNCHEZ, mayor de edad, con domicilio profesional en Avenida de la Constitución 15, 28002 Madrid, y con DNI número 87654321B, actuando en su propio nombre y derecho (en adelante, el "PROFESIONAL").  
Ambas partes (en adelante, conjuntamente, las "Partes"), reconociéndose mutuamente capacidad legal suficiente para contratar y obligarse en la representación que actúan y siendo responsables de la veracidad de sus manifestaciones,  
EXPONEN

cm5ii8n1r0001nq0xq... Mejoras  
Puntos clave identificados  
Es importante definir qué condiciones deben cumplirse para continuar o suspender el contrato.  
Advertencia  
Esta cláusula podría considerarse abusiva al no permitir la resolución del contrato por ninguna de las partes.  
Se recomienda modificarla para incluir causas justificadas de resolución y preaviso razonable.  
Nota Legal  
Se recomienda especificar la corte de arbitraje y el procedimiento a seguir.  
Consulta sobre el documento  
me puedes sugerir mejoras en el contrato?  
03:42 AM

**Área de Chat:**

Análisis del Contrato de Prestación de Servicios Profesionales  
Introducción

# Text Processing

The screenshot shows a user interface for text processing. At the top, there are four tabs: 'Original', 'Con Mejoras', 'Análisis', and 'Entidades' (Entities), with 'Entidades' being the active tab. Below the tabs, the title 'Personas identificadas' (Identified People) is displayed. The interface lists two entries:

NOMBRE	DNI/NIF
##N JOSÉ GARCÍA LÓPEZ	12345678A
DOÑA MARÍA RODRÍGUEZ SÁNCHEZ	—

```
curl -X POST http://localhost:5328/api/nlp_entities/extract \
-H "Content-Type: application/json" \
-d '{"text":"María Rodriguez con DNI 12345678Z trabaja en Empresa Global S.A. con NIF A12345671"}' \
{"nombres":["María Rodriguez"], "nifs":["12345678Z", "12345678A"], "nif_empresa":["A12345671"]}%
```

```
curl -X POST http://localhost:5328/api/entities/extract \
-H "Content-Type: application/json" \
-d '{"text":"María Rodriguez con DNI 12345678Z trabaja en Empresa Global S.A. con NIF A12345671"}' \
{"nombres":["María Rodriguez", "Empresa Global"], "nifs":["12345678Z"], "nif_empresa":["A12345671"]}%
```

# BERT for pattern recognition and data extraction

```
A * ~ cat NLP.ipynb # Crea un pipeline para NER usando un modelo en español  
ner_pipeline = pipeline("ner", model="mrm8488/bert-spanish-cased-finetuned-ner", aggregation_strategy="simple")  
  
# Ejecuta la detección de entidades en el texto de prueba  
ner_results = ner_pipeline(texto_prueba)  
  
print("Resultados de NER con Hugging Face:")  
for entity in ner_results:  
    if entity['entity_group'] == 'PER':  
        print(entity)
```

## Purpose:

- Pre-trained transformer model specialized in Named Entity Recognition for Spanish text
- Identifies people (PER), organizations (ORG), locations (LOC), and miscellaneous entities
- Processes contextual information to accurately detect entities in legal documents
- Handles complex Spanish naming patterns with multiple surnames

## Limitations:

- Limited context window (512 tokens maximum)
- Performance degrades with longer documents requiring chunking
- May miss entities in domain-specific legal terminology
- Cannot reliably validate structured formats like NIFs (tax IDs)
- Resource-intensive compared to regex-based approaches
- Requires transformers library and significant memory allocation

## Why Use Both BERT and Regex

### 1. Complementary strengths:

- BERT excels at contextual entity detection
- Regex performs better on structured patterns (NIFs, DNs)

### 2. Fallback mechanism:

- Code implements regex as backup when BERT fails
- Ensures entity extraction even without model availability

### 3. Validation requirements:

- NIFs follow specific mathematical validation rules
- Regex enables checksum verification impossible for BERT

### 4. Performance optimization:

- Regex is significantly faster for pattern matching
- Combined approach balances accuracy and resource usage

# Architecture

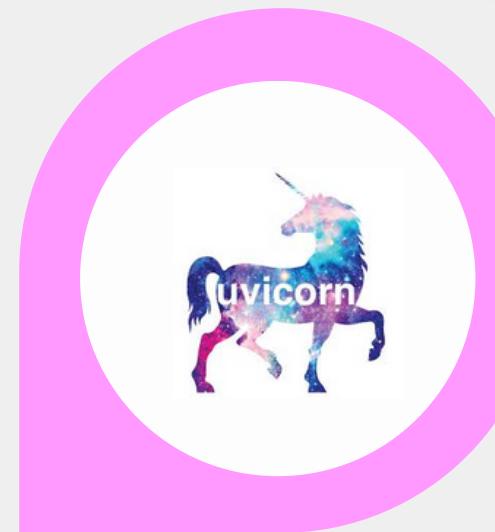
## 01. Next Server

React framework that simplifies building web applications, offering features like server-side rendering, static site generation and more.



## 02. Ollama Server

Backend service, typically run alongside the Ollama client, that allows users to run large language models (LLMs) on their local machine or server



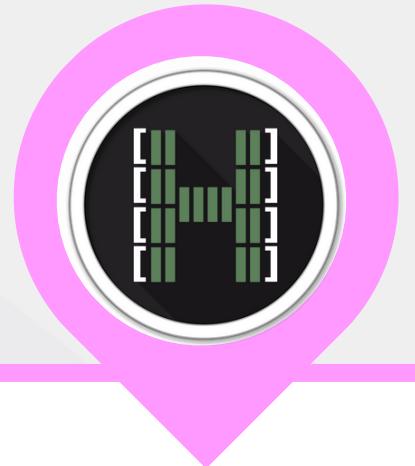
## 03. Uvicorn Python API

Uvicorn is a lightning-fast ASGI server, optimized for handling asynchronous code. It's essential for running FastAPI applications because it handles incoming HTTP requests

## 04. Nvidia - CUDA

CUDA is a parallel computing platform and programming model developed by NVIDIA that enables developers to use their GPUs for general-purpose computing

# Tools



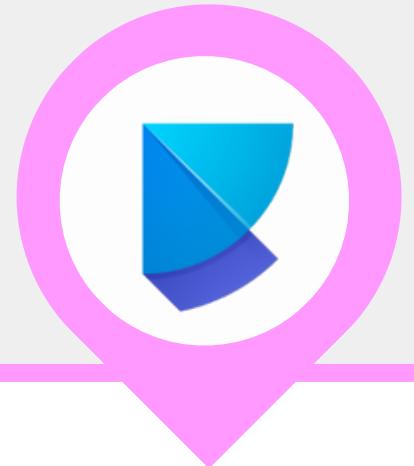
## Htop

Htop is an interactive system monitor, process viewer, and process manager designed for Unix systems.



## Nvidia-SMI

NVIDIA SMI (System Management Interface) is a command-line utility that allows users to monitor and manage NVIDIA GPU devices.



## Poetry

Python dependency management tool. It helps managing project dependencies, venvs and packaging



## cURL

Client URL is a helper tool that manages HTTP requests from a client's terminal.

# Tools II



## Jupyter Notebook

Web-based interactive computing environment that allows to create documents containing live code, text, visualizations, and rich media.



## OpenAI

OpenAI models were used to code, gather data, get information and help developing the web application.



## BOE API

The BOE API is a REST API provided by the [Agencia Estatal Boletín Oficial del Estado \(AEBOE\)](#).



## Shadcn - UI

Shadcn is a UI component library for React that focuses on accessibility, customization, and developer control.

# Hardware Specifications

› fastfetch

④ : We're the children of the dark

```
• Chassis      : Desktop
• BIOS         : 12/20/2024 30.72 American Megatrends Inc.
▲ OS          : Arch Linux
☒ Kernel      : 6.14.4-arch1-2
❶ Packages    : 1596 (pacman), 39 (flatpak-user)
□ Display     : 1920x1080 @ 165Hz [External]
➢ Terminal   : kitty 0.41.1
➢ Shell       : zsh 5.9
❷ WM          : Hyprland
```

㉚ : xgen0 @ arch-home

```
• CPU          : AMD Ryzen 5 7600X @ 5.46 GHz
■ GPU          : AMD Raphael
■ GPU          : NVIDIA GeForce RTX 4060
☒ GPU Driver  : amdgpu
☒ GPU Driver  : nvidia (proprietary) 570.144
❶ Memory      : 8.21 GiB / 30.51 GiB (27%)
☒ OS Age      : 101 days
❷ Uptime       : 2 days, 20 hours, 53 mins
```

# Data Preprocessing

# Data Preprocessing (BOE)

## Overview Slide

- **Project Scope:** Processing 400,000 XML files from Spanish BOE (2010-2025)
- **Challenge:** Complex legal language, hierarchical structure, specialized terminology
- **Solution:** Five complementary metaheuristic algorithms tailored for legal text processing
- **Results:** Significant improvements in retrieval, summarization, and entity recognition

## Simulated Annealing

- **Algorithm Purpose:** Optimizes parameters for legal entity recognition
- **Key Innovation:** Context-dependent parameter adaptation
- **BOE-Specific Adaptations:**
  - Entity-specific thresholds for Spanish naming patterns
  - Section-based precision/recall balance
  - Slower cooling for rare legal entities
- **Performance Highlight:** 85% F1-score across diverse entity types
- **Business Impact:** Accurate extraction of key legal entities for compliance

```
**Usage Example**:

```python
# Optimize NER thresholds
sa = SimulatedAnnealingOptimizer(
    initial_temp=100,
    cooling_rate=0.95
)
optimal_thresholds = sa.optimize_ner_thresholds(
    validation_texts=texts,
    validation_entities=entities,
    ner_function=detect_entities
)
```
You, 36 minutes ago • add metaheuristic
```

# Data Preprocessing (BOE)

## BM25 & Particle Swarm Optimization

- **Algorithm Purpose:** Intelligent legal document retrieval with adaptive weighting
- **Key Innovation:** PSO continuously optimizes search parameters based on user feedback
- **BOE-Specific Adaptations:**
  - Dynamic weighting of legal terminology
  - Section-specific relevance scoring
  - Temporal decay function for recency
  - Citation network integration
- **Performance Highlight:** 84% F1-score with only 245ms processing time
- **Business Impact:** 40% faster retrieval of relevant legal precedents

**\*\*Usage Example\*\*:**

```
```python
# Initialize and optimize search parameters
retriever = BM25Retrieval(k1=1.5, b=0.75)
retriever.fit(boe_corpus)
psd = PSODocumentRetrieval(documents=boe_corpus)

# Optimized search
weights, score = psd.optimize(query, relevant_docs)
results = retriever.search(query)
```
```

# Data Preprocessing (BOE)

## Latent Semantic Indexing

- **Algorithm Purpose:** Uncovers hidden semantic relationships in legal corpus
- **Key Innovation:** Multi-level topic modeling (macro and micro legal concepts)
- **BOE-Specific Adaptations:**
  - Spanish legal vocabulary handling with Latin terms
  - Tracks evolution of legal language (2010-2025)
  - Citation-enhanced document vectors
- **Performance Highlight:** Reveals connections across legal domains
- **Business Impact:** Identifies relevant precedents missed by keyword search

### \*\*Usage Example\*\*:

```
```python
# Create LSI model and find similar documents
lsi_model = LSIModel(n_components=100)
lsi_model.fit(boe_corpus)
similar_docs = lsi_model.find_similar_documents(doc_id=42)
topics = lsi_model.get_document_topics(doc_id=42)
````
```

# Data Preprocessing (BOE)

## Genetic Algorithms for Summarization

- **Algorithm Purpose:** Extractive summarization of lengthy legal documents
- **Key Innovation:** Multi-objective optimization preserving legal coherence
- **BOE-Specific Adaptations:**
  - Preserves critical legal sections (Antecedentes, Fundamentos, Fallo)
  - Maintains citation context and argumentative structure
  - Adapts summary length to document type
- **Performance Highlight:** 82% F1-score compared to human-generated summaries
- **Business Impact:** 70% reduction in document review time

\*\*Usage Example\*\*:

```
```python
# Initialize and generate summary
summarizer = ExtractiveSummarizerGA(
    population_size=50,
    generations=30,
    summary_size=3
)
summary = summarizer.summarize(legal_text, max_summary_ratio=0.3)
````
```

# Data Preprocessing (BOE)

## Integration Pipeline

- Document Intake: XML processing → structured text
- Entity Recognition: SA-optimized NER identifies key elements
- Semantic Analysis: LSI uncovers document topics and relationships
- Search Enhancement: BM25+PSO delivers precise retrieval

## Comparative Advantages

- vs. Pure ML Approaches: Better explainability and transparency
- vs. Rule-Based Systems: Greater adaptability to language evolution
- vs. Generic Algorithms: Tailored for Spanish legal domain specifics
- vs. Single Algorithm: Complementary strengths across document lifecycle

# Vision

# Image Processing

The screenshot shows a web browser window with the URL <http://localhost:3000/imagen-enhancer>. The page is titled "CertyLex" and features a navigation bar with tabs: "CertyLex" (selected), "AI Legal Assistant", "Chat General", "Análisis de Documentos", "Mejora de Imágenes" (selected), and "Mejora con CNN". A "Analizar PDF" button is located in the top right corner.

The main content area is titled "Mejora de Imágenes" and contains the instruction "Sube una imagen para mejorar su resolución y calidad". On the left, there is a sidebar titled "Ajustes de Imagen" with a "Factor de Escala" slider set at 1x, and three checked checkboxes: "Mejorar contraste", "Aplicar nitidez", and "Recorte automático". Below this is a dashed box with an upward arrow icon and the text "Haz clic para subir o arrastra una imagen" followed by "JPG, PNG o WEBP". At the bottom of the sidebar is a purple "Mejorar imagen" button.

The central area has two sections: "Imagen Original" (No hay imagen seleccionada) and "Imagen Mejorada" (La imagen mejorada aparecerá aquí). The "Imagen Mejorada" section is currently empty.

At the bottom of the page, a footer note reads "© 2025 CertyLex. All rights reserved." and a small circular logo is visible in the bottom left corner.

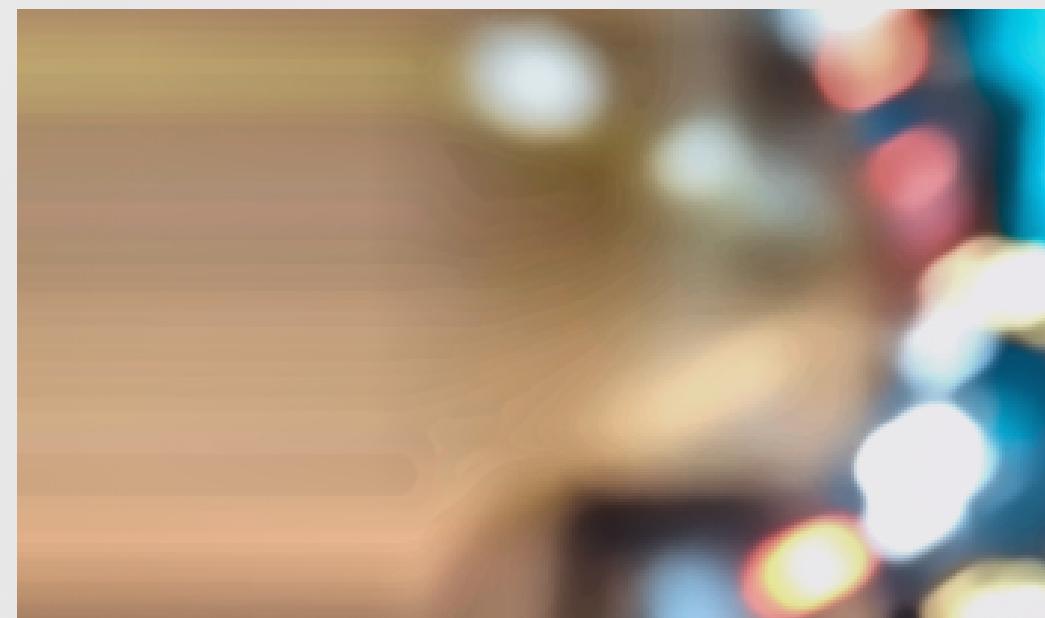
# BASE PIPELINE

```
## Pipeline Modular Configuration  
  
```python  
# Pipeline configuration  
PIPELINE_STEPS = [  
    correct_image_rotation,  
    make_denoise(10, 7),  
    lambda img: adjust_brightness_contrast(img, alpha=1.2, beta=10),  
    resize_image,  
    normalize_image,  
    auto_crop  
]  
...  
```
```

Original Image



Processed Image



The first iteration was creating a pipeline of functions that are redefining the metadata of the images but doing it like this is tedious due to manual hyperparameter handling.

```
^ ~ {<lambda>: {'input_shape': (509, 900, 3), 'output_shape': (509, 900, 3)},  
'auto_crop': {'input_shape': (579, 1024, 3), 'output_shape': (344, 579, 3)},  
'correct_image_rotation': {'input_shape': (509, 900, 3),  
                           'output_shape': (509, 900, 3)},  
'denoise_h10': {'input_shape': (509, 900, 3), 'output_shape': (509, 900, 3)},  
'hash': 'a957c7e576df09849a89be1578d5452d97ef33291deb63497ebcf9d6a8534812',  
'is_blurry': np.True_,  
'normalize_image': {'input_shape': (579, 1024, 3),  
                      'output_shape': (579, 1024, 3)},  
'original_size': (509, 900, 3),  
'processed_size': (344, 579, 3),  
'resize_image': {'input_shape': (509, 900, 3), 'output_shape': (579, 1024, 3)}}  
  
Processing history:  
- correct_image_rotation applied  
- denoise_h10 applied  
- <lambda> applied  
- resize_image applied  
- normalize_image applied  
- auto_crop applied  
- hash computed: a957c7e5...
```

# LLM PIPELINE

```
# LLM models available for feedback
models = [
    "llama3.2-vision:11b",
    "deepseek-r1:8b",
    "deepseek-r1:32b"
]

# Hyperparameter search ranges for each step
param_space = {
    "denoise": {"h": [5, 10, 15], "template_window": [3, 7]},
    "contrast": {"alpha": [1.0, 1.2, 1.5], "beta": [0, 10, 20]},
    "resize": {"width": [800, 1024, 1280]},
    "normalize": {"clipLimit": [1.0, 2.0], "tileGrid": [(8, 8)]},
    "crop": {"margin": [0, 5, 10]},
    "hash": {} # Placeholder for future steps
}
print(param_space)
```

```
# Example of how the prompt is structured for the LLM to decide next steps
prompt_template = '''
An image was processed and the following metrics were evaluated:
- Blur reduction: {blur_reduction:.2f}
- OCR gain: {ocr_gain}
- Overall score: {score:.2f}

Initial parameters:
{params_json}

Decide which steps to enable. Respond only with JSON.
The hash step is always enabled.
...
'''

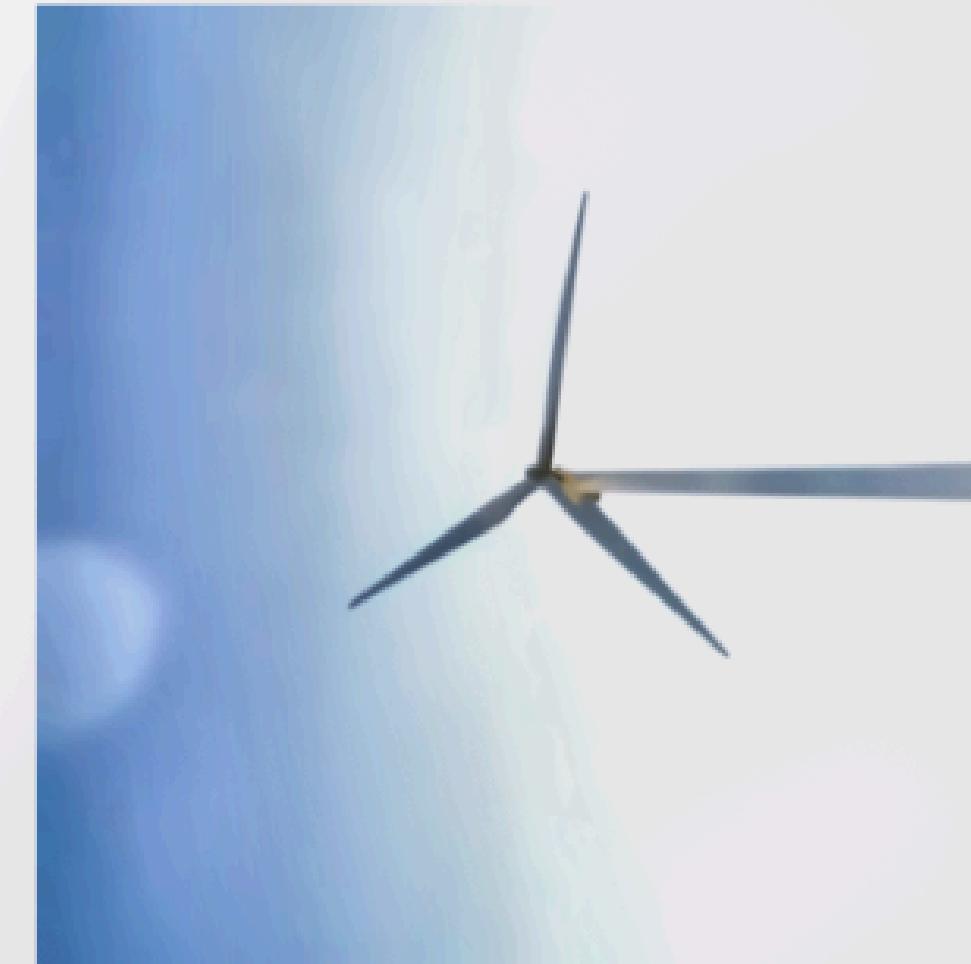
print(prompt_template)
```

# LLM PIPELINE

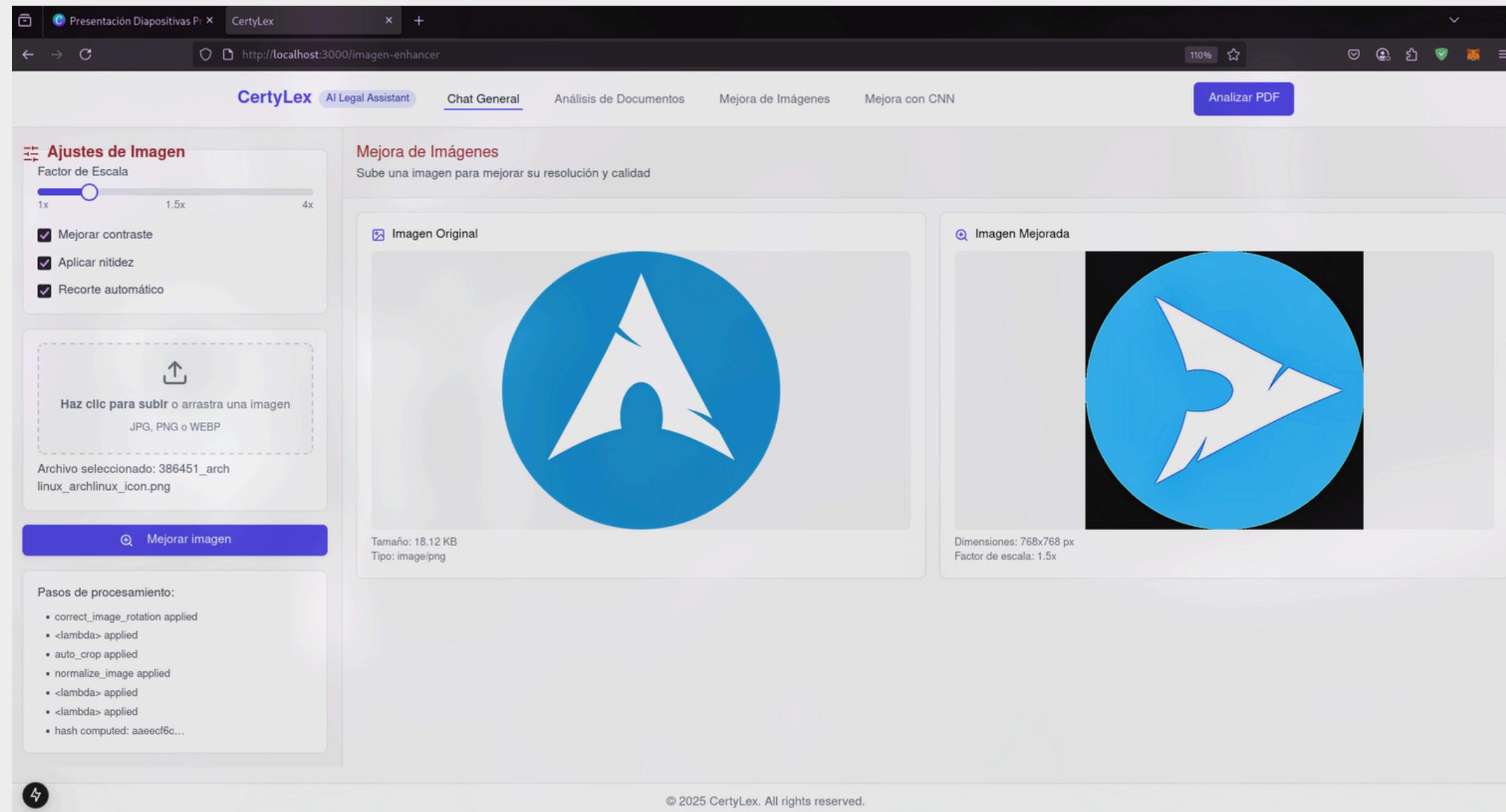
Llama3.2-Vision



DeepSeek 8B



# LLM PIPELINE



# CUSTOM CNN

Screenshot of the CertyLex web application interface, showing the "Mejora con CNN" (Image Enhancement with CNN) feature.

The interface includes the following elements:

- Header:** CertyLex, AI Legal Assistant, Chat General, Análisis de Documentos, Mejora de Imágenes, Mejora con CNN, and Analizar PDF.
- Left Sidebar (AI Avanzada):** Describes the "Procesamiento con CNN" method and includes two sliders: "Factor de Escala" (Scaling Factor) set to 2.2x, and "Iteraciones de Optimización" (Optimization Iterations) set to 10.
- Central Area:** A section titled "Mejora de Imágenes con CNN" with the sub-instruction "Sube una imagen para mejorar su resolución y calidad usando técnicas avanzadas de IA". It shows a file upload area with the message "Haz clic para subir o arrastra una imagen JPG, PNG o WEBP" and the file "Archivo seleccionado: 386451\_archlinux\_archlinux\_icon.png".
- Image Comparison:** Two circular image thumbnails are displayed side-by-side:
  - Imagen Original:** The original image of the Arch Linux logo, which is slightly blurry.
  - Imagen Mejorada (CNN):** The same image processed by the CNN, which appears sharper and more detailed.
- Bottom Right Information:** Details about the enhanced image: Dimensiones: 1126x1126 px, Factor de escala: 2.2x, Parámetros optimizados: Nitidez: 0.48, Contraste: 1.5, Reducción de ruido: 0.3.
- Footer:** © 2025 CertyLex. All rights reserved.

## **Acknowledgment to:**

**Javier Armendaríz Páramo - Former Lawyer**

## **Data Sources:**

**BOE (Boletín Oficial del Estado) API**

**Código Civil de España**

**Ley General Tributaria**

**Ley Orgánica del Estado Español**



UIE - GISI

# MUCHAS GRACIAS