# —ANEXOS—

# USO DE INTELIGENCIA ARTIFICIAL PARA LA CARACTERIZACIÓN DE LA CARGA REGULATORIA AMBIENTAL DE PROYECTOS EN DESARROLLO

Comisión Nacional de Evaluación y Productividad

Esta versión: 2025-10-20

# Índice

1 Anexo 1: Flujo de Trabajo para la Extracción de Obligaciones desde Resoluciones de Calificación Ambiental 6

2 Anexo 2: Validación de los Resultados Obtenidos por IA 21

## Listado de Figuras


1 Flujo de Trabajo . . . . . 7
2 Comparación de resultados de ambas metodologías para una submuestra de RCAs . . . . . 9


## Listado de Tablas


3 Cálculo de Tamaño Muestral Mínimo . . . . . 23
4 Composición de la Muestra de RCAs . . . . . 23
5 Criterio de Extracción de Obligaciones por Sección . . . . . 24
6 Resumen de Resultados de Validación de la Clasificación . . . . . 31

## Acrónimos

**CMN** Consejo de Monumentos Nacionales

**CNEP** Comisión Nacional de Evaluación y Productividad

**DFL** Decreto con Fuerza de Ley

**DGA** Dirección General de Aguas

**DIA** Declaración de Impacto Ambiental

**DS** Decreto Supremo

**EIA** Estudio de Impacto Ambiental

**IA** Inteligencia Artificial

**IA Gen** Inteligencia Artificial Generativa

**ICE** Informe Consolidado de Evaluación

**IFC** Informe Favorable para la Construcción

**LGUC** Ley General de Urbanismo y Construcciones

**LLM** Large Language Model (Modelo de Lenguaje Extenso)

**MINSAL** Ministerio de Salud

**MMA** Ministerio del Medio Ambiente

**OCR** Optical Character Recognition (Reconocimiento Óptico de Caracteres)

**OGUC** Ordenanza de Urbanismo y Construcciones

**PAS** Permiso Ambiental Sectorial

**RCA** Resolución de Calificación Ambiental

**RESPEL** Residuos Peligrosos

**RETC** Registro de Emisiones y Transferencias de Contaminantes

**RSEIA** Reglamento del Sistema de Evaluación de Impacto Ambiental

**SAG** Servicio Agrícola y Ganadero

**SEA** Servicio de Evaluación Ambiental

**SEIA** Sistema de Evaluación de Impacto Ambiental

**SMA** Superintendencia del Medio Ambiente

**SNGM** Servicio Nacional de Geología y Minería

# 1. Anexo 1: Flujo de Trabajo para la Extracción de Obligaciones desde Resoluciones de Calificación Ambiental

El presente anexo detalla la metodología de extracción y calificación de obligaciones ambientales a partir de las Resoluciones de Calificación Ambiental (RCA) emitidas por el Servicio de Evaluación Ambiental (SEA) de Chile utilizando Inteligencia Artificial.

El procedimiento desarrollado se articula en dos tareas principales, cada una con una lógica distinta:

- En el primer paso, el desafío principal consiste en extraer de manera precisa la información desde los archivos PDF, muchos de los cuales se encuentran en formatos de baja calidad, con estructuras heterogéneas o escaneados como imágenes. El objetivo aquí es copiar y pegar la información de forma fidedigna, generando una base de datos intermedia que respete la numeración, el orden y la estructura original de los documentos.
- En el segundo paso, el foco cambia: ya no se trata de extraer texto, sino de interpretar qué constituye una obligación ambiental. Para ello, se emplean modelos de lenguaje (LLMs) entrenados con instrucciones y ejemplos proporcionados por expertos, con el fin de identificar, clasificar y estandarizar las obligaciones a partir de los registros extraídos en la etapa anterior.

La Figura 1 resume gráficamente estos dos pasos fundamentales del proceso.

## Figura 1: Flujo de Trabajo

| | |
|---|---|
| Referencia al ICE para mayores detalles | Ver numeral 8.1 del ICE del proyecto. |

| 7.1.2. Sistema de pulverización de la planta clasificadora de muros. | |
|---|---|
| Tipo de medida | Mitigación. |
| Componente(s) ambiental(es) objeto de protección | Riesgo para la salud de la población. |
| Impacto asociado | Calidad del aire. |
| Objetivo, descripción y justificación | Objetivo: Minimizar las emisiones de material particulado por clasificación y proceso de chancado para material de muros del DR.<br><br>Descripción: Incorporar y mantener sistema de pulverización en los equipos de la planta clasificadora de muros, además de humectar las áreas de trabajo cuando las condiciones ambientales lo requieran.<br><br>Justificación: Minimizar las emisiones de material particulado por clasificación y proceso de chancado para material de muros de DR. |
| Lugar, forma y oportunidad de implementación | Lugar: En la planta clasificadora para material de muros del DR.<br><br>Forma: Incorporar y mantener sistema de pulverización en los equipos de la planta clasificadora de muros.<br><br>Oportunidad: Durante las fases de construcción y operación, mientras la planta clasificadora se encuentre operativa. |

**Paso 1**

| 0 | 1 |
|---|---|
| 7.1.2. Sistema de pulverización | de la planta clasificadora de muros. |
| Tipo de medida | Mitigación. |
| Componente(s) ambiental(es) objeto de protección | Riesgo para la salud de la población. |
| Impacto asociado | Calidad del aire. |
| Objetivo, descripción y justificación | Objetivo: Minimizar las emisiones de material particulado por clasificación y proceso de chancado para material de muros del DR. Descripción: Incorporar … |
| Lugar, forma y oportunidad de implementación | Lugar: En la planta clasificadora para material de muros del DR. Forma: Incorporar y mantener sistema de pulverización en los equipos de la planta clasificadora de muros. …… |

**Primer Paso:** Cada subconsiderando se convierte en una fila, manteniendo su numeración para identificar su origen y revisión.

| 0 | 1 |
|---|---|
| 7.1.2. Sistema de pulverización | de la planta clasificadora de muros. |
| Tipo de medida | Mitigación. |
| Componente(s) ambiental(es) objeto de protección | Riesgo para la salud de la población. |
| Impacto asociado | Calidad del aire. |
| Objetivo, descripción y justificación | Objetivo: Minimizar las emisiones de material particulado por clasificación y proceso de chancado para material de muros del DR. Descripción: Incorporar … |
| Lugar, forma y oportunidad de implementación | Lugar: En la planta clasificadora para material de muros del DR. Forma: Incorporar y mantener sistema de pulverización en los equipos de la planta clasificadora de muros. …… |

**Paso 2**

| seccion | numero_tabla | resumen | componente_materia |
|---|---|---|---|
| pas | 6.2.1 | Se indica la | residuos sólidos |
| pas | 6.2.2 | Se establec | sustancias peligrosas |
| pas | 6.2.3 | Se solicita p | social |
| forma_cum | 7.1 | Implement | aire |
| forma_cum | 7.1 | Prohibir la | aire |
| forma_cum | 7.1 | Humectar c | aire |
| forma_cum | 7.1 | Cubrir cam | aire |
| forma_cum | 7.1 | Exigir que t | aire |

**Segundo Paso:** Extraer múltiples obligaciones individuales de cada fila y clasificarlas en categorías predefinidas.

Se implementan dos metodologías en paralelo, desarrolladas por equipos distintos. La decisión de avanzar por dos vías responde a múltiples propósitos:

- Explorar herramientas alternativas y evaluar sus ventajas y limitaciones.
- Corroborar si, a través de enfoques diferentes, se logran resultados consistentes.
- Identificar en qué tipos de documentos o tareas específicas un método puede superar al otro.

Las diferencias fundamentales entre ambas metodologías se encuentran en el primer paso:

- La **Metodología Alternativa** se apoyó en herramientas pagadas de *document intelligence*[1] —en particular *Amazon Textract*— para extraer textos y tablas de manera estructurada directamente desde los PDF.
- La **Metodología Principal** utilizó herramientas gratuitas —de menor calidad— para el reconocimiento de tablas y OCR, pero compensando esta limitación mediante el uso de LLMs para la extracción de la información.

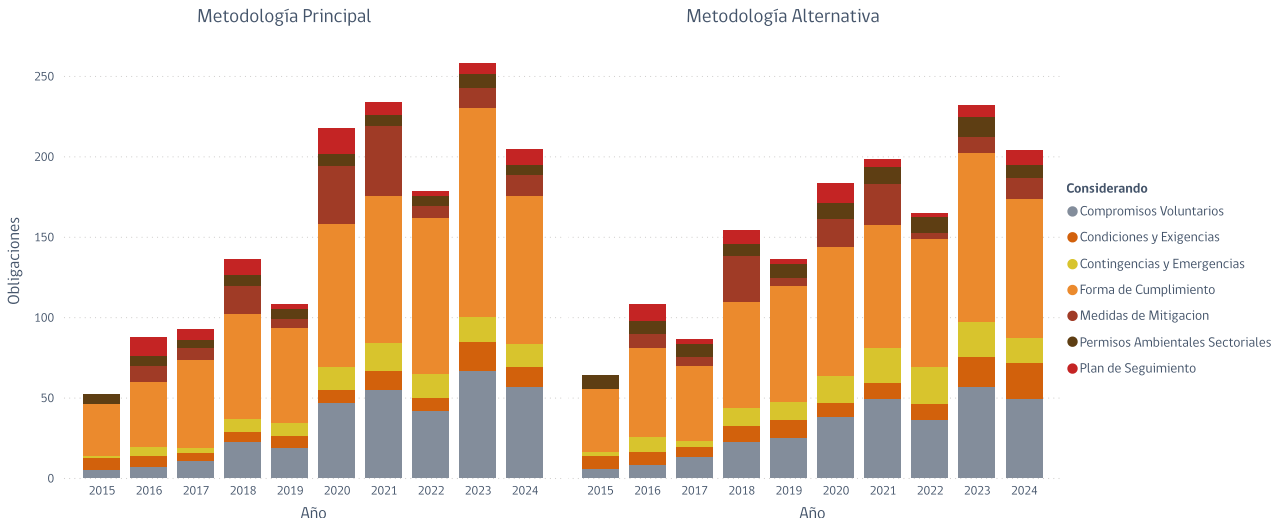
En la fase interpretativa, ambas metodologías convergen en el uso del mismo modelo, pero difieren en el insumo que procesan:

- En la **Metodología Alternativa**, los bloques provenientes de *Amazon Textract* se envían directamente al modelo para su interpretación.
- En la **Metodología Principal**, antes de llegar al LLM, la información ya ha sido organizada en una base intermedia con categorías predefinidas, lo que facilita el análisis.

Tal como se observa en la Figura 2, los resultados agregados de ambas metodologías son similares:

[1] *Document intelligence* se refiere a tecnologías que superan al OCR tradicional al no solo digitalizar texto, sino también interpretar la estructura del documento como un todo. A diferencia de un LLM, su función principal no es el procesamiento de lenguaje natural, sino la extracción de datos basada en la composición visual del archivo.

**Figura 2: Comparación de resultados de ambas metodologías para una submuestra de RCAs**




Fuente: Elaboración Propia

Vale la pena resaltar que en ciertos casos, la metodología principal ofrecía un mayor nivel de granularidad que la alternativa, mientras que en otros sucedía lo contrario. Esto pone de manifiesto que la delimitación de las obligaciones ambientales no responde a un criterio único y definitivo, sino que involucra márgenes de interpretación, dependientes del contexto y de las particularidades de cada RCA, y que con múltiples iteraciones fueron convergiendo hacia los resultados deseados.

Finalmente, se optó por utilizar la Metodología Principal como base del estudio, al mostrar un mejor desempeño en la práctica. No obstante, el desarrollo de dos enfoques en paralelo fue altamente valioso: permitió validar la robustez de los resultados, comprender mejor los desafíos del problema y obtener aprendizajes relevantes para futuros desarrollos en el uso de herramientas de IA Gen.

La implementación de esta metodología está disponible como código abierto en el repositorio RCAs-CNEP

## Preparación de datos de ambas metodologías

- Se ingresa al portal de búsqueda de proyectos del Servicio de Evaluación Ambiental (SEIA) en este enlace, usando los siguientes filtros para obtener los proyectos de interés:
  - **Estado del proyecto:** Aprobados.
  - **Fecha de Calificación:** Entre el 1 de enero de 2015 y el 31 de diciembre de 2024.
  - **Sector productivo:** Minería y Energía.
  - **Tipología del proyecto (DS 40):** i.1), i.3), b1), c).

  Al aplicar estos filtros, se genera un archivo excel en el que cada fila corresponde a una Resolución de Calificación Ambiental (RCA) distinta, y que contiene las siguientes columnas:

- Nombre
- WEB (URL de descarga)
- Tipo
- Región
- Comunas
- Provincias
- Tipología
- Razón de Ingreso
- Titular
- Inversión (MMU$)
- Fecha presentación
- Estado
- Fecha calificación
- Sector productivo
- Latitud punto representativo
- Longitud punto representativo

- **Descarga de Archivos PDF** Se utilizan herramientas estándar de *webscraping* (librería *selenium*) para automatizar la interacción con el sitio web y asegurar la descarga correcta de cada archivo presente en el listado, manejando botones y cuadros emergentes cuando sea necesario.
- **Extracción de Textos de Archivos Escaneados** Dado que, una parte importante de los archivos son imágenes escaneadas y no permiten la extracción de tablas y textos directamente, se aplica un procedimiento adicional para estos documentos. Particularmente, se emplea la herramienta *docTR* para realizar Reconocimiento Óptico de Caracteres (OCR), lo que convierte las imágenes de textos escaneados en texto editable sin costo.

## Metodología Principal

### Paso 1: Extracción de Subconsiderandos de los Archivos

Los archivos suelen estar estructurados en subconsiderandos enumerados (7.1.1, 7.1.2, 7.2.1, etc), generalmente cada subconsiderando es una tabla, pero no siempre es así. Este paso convierte cada subconsiderando de cada considerando de interés en un registro único del archivo intermedio, con columnas predefinidas, independientemente de su estructura en el archivo de origen.

**Considerandos de Interés:**

- Medidas de Mitigación
- Planes de Seguimiento
- Permisos Ambientales Sectoriales
- Forma de Cumplimiento
- Compromisos Voluntarios
- Condiciones y Exigencias
- Contingencias y Emergencias

**Procedimiento Realizado para cada documento:**

1. **Extracción del Texto**
2. **Identificación de Considerandos:** Mediante delimitadores específicos, se identifican los considerandos relevantes en el texto, los cuáles pueden presentarse en distinto orden, utilizando múltiples delimitadores[2] específicos posibles para cada considerando, y admitiendo diferencias leves en su redacción[3]:

| Considerando | Delimitador de Ejemplo |
|---|---|
| Vistos | VISTOS |
| Considerando | CONSIDERANDO |
| Mitigación | Que, del proceso de evaluación de impacto ambiental del Proyecto puede concluirse que las siguientes medidas de mitigación... |
| PAS | Que resultan aplicables al Proyecto los siguientes permisos ambientales sectoriales, asociados a las correspondientes... |
| Forma de Cumplimiento | Que, de acuerdo a los antecedentes que constan en el expediente de evaluación, la forma de cumplimiento de la normativa... |
| Condiciones y Exigencias | Que, para ejecutar el Proyecto deben cumplirse las siguientes condiciones o exigencias, en concordancia con el artículo 25... |
| Compromisos Voluntarios | Que, durante el procedimiento de evaluación de la DIA se establecieron los siguientes compromisos ambientales voluntarios... |
| Contingencias y Emergencias | Que, las medidas relevantes del Plan de Prevención de Contingencias y del Plan de Emergencias, son las siguientes... |

[2] Que pueden contener espacios y otros caracteres entremedio.

[3] Se usa *fuzzywuzzy* para admitir delimitadores que tengan cercanía textual.

| Considerando | Delimitador de Ejemplo |
|---|---|
| Considerandos Otros | En la presente evaluación no se realizaron reuniones con grupos humanos pertenecientes a pueblos indígenas del artículo 86... |
| Resuelvo | RESUELVO |

Fuente: Elaboración Propia

3. **División en Chunks (Fragmentos):** Si un considerando supera los 6.500 *tokens*[4] de contenido, se divide en fragmentos o *chunks*[5] de máximo 4.500 *tokens*, con un solapamiento de 1.000 *tokens* para evitar pérdida de información.

4. **Extracción de Tablas (solo para archivos no escaneados):** Adicional al texto, con la finalidad de que el LLM tenga mayor comprensión de la estructura de tablas, se extraen las tablas presentes en cada documento (buscando sólo dentro del texto de los considerandos relevantes) utilizando la herramienta *tabula* de lectura de tablas PDF. Utilizando distintas funciones se reconoce el número de tabla, se unen tablas que debiesen lógicamente estar juntas, pero que *tabula* toma de manera separada[6], y se asigna cada atributo encontrado en la tabla a alguno de los siguientes campos posibles, admitiendo diferencias leves en la redacción[7]:

| Campo | Texto de Ejemplo (Pueden presentarse de muchas maneras distintas) |
|---|---|
| componente_materia | componente/materia |
| nombre_compromiso | nombre del compromiso |
| riesgo_contingencia | riesgo o contingencia |
| normas | normas legales |
| fase | fases |
| parte_asociada | parte, obra o acción a la que aplica |
| lugar_forma_oportunidad | lugar, forma y oportunidad de implementación |
| forma_cumplimiento | forma de cumplimiento |
| tipo_medida | tipo de medida |
| indicador_cumplimiento | indicador que acredita su cumplimiento |
| control_seguimiento | forma de control y seguimiento |
| ice | referencia al ice para mayor detalle |

[4] Es la unidad fundamental de texto que un modelo de lenguaje procesa. Puede ser una palabra, parte de una palabra, una combinación de palabras, o un signo de puntuación.

[5] Es un fragmento o bloque de texto, compuesto por un conjunto de *tokens*, que se crea al dividir un documento más extenso. Esta segmentación es necesaria para procesar textos que exceden el límite de *tokens* (la ventana de contexto) que un modelo puede analizar de una sola vez.

[6] Por ejemplo, por un salto de página

[7] Buscando sin espacios ni caracteres especiales

| Campo | Texto de Ejemplo (Pueden presentarse de muchas maneras distintas) |
|---|---|
| otros_legal | otros cuerpos legales |
| impacto | impacto asociado |
| objetivo_descripcion_justificacion | objetivo, descripción y justificación |
| calificacion | calificación de la parte u obra |
| condiciones | condiciones o exigencias específicas del pronunciamiento |
| pronunciamiento | pronunciamiento del órgano competente |
| acciones | acciones o medidas a implementar para prevenir la contingencia |
| referencia | referencia a documentos del expediente de evaluación |
| sma | oportunidad y vías de comunicación a la sma... |

Fuente: Elaboración Propia

5. **Asignación de Tablas a Considerandos:** Se asigna cada tabla a su considerando correspondiente en base a las variables presentes en dicha tabla[8] y el listado de considerandos identificados en el documento analizado y su orden.

6. **Procesamiento con Modelo de Lenguaje:** Para cada combinación de *chunks* y conjunto de tablas, se utiliza el modelo *gpt-4o-mini* de OpenAI. Esta función tiene como argumentos:

   *a)* **Texto de Entrada**: se proporciona todo el texto del *chunk* y además de las tablas identificadas, explicitando al modelo: *"Para facilitar tu trabajo, también proporcionaré una tabla JSON (la cual puede contener errores y puede No coincidir exactamente con el texto, sólo es una guía)"*. En caso de que no se hayan identificado tablas entonces sólo se entrega el texto del *chunk*.

   *b)* **Instrucciones**: que son diferentes para cada uno de los considerandos relevantes, ya que en cada considerando son diferentes los atributos de interés. Todas las instrucciones siguen esta base:

```
## Contexto:
Te proporcionaré el texto de ...(se introduce el considerando)... de una
    Resolución de Calificación Ambiental (RCA) emitida por el Servicio de
    Evaluación Ambiental de Chile para distintos proyectos.
El texto se dividirá siempre en tablas que contendrán diversas variables.

## Objetivo:
Tu tarea es registrar tal como están en el texto original ciertos elementos de
     cada tabla (que pueden estar expresados de maneras ligeramente diferentes
    ):
```

[8] Por ejemplo, si una tabla presenta el atributo "condiciones" entonces se asigna automáticamente al considerando 'PAS', porque es la única con dicho atributo.

```
...
(detalles específicos del considerando)
...
Debes entregar absolutamente todo el texto completo asociado a cada campo
    buscado, sin excepción, por más extenso que sea.

Debes asegurar que todas las variables tengan textos legibles:
- Espacios entre palabras correctamente asignados.
- Corregir errores ortográficos.
- Eliminar textos extraños como "@$$#%!opw¿".

- Recuerda que la instrucción más relevante es: si el objetivo es entregar el
    texto completo entonces debes entregar absolutamente todo el texto
    completo asociado a dicho campo buscado, sin excepción. Si ves anexos o NO
     estas seguro de si dicho texto forma parte o no del campo, incluyelos
    igualmente.
- El texto "Para validar las firmas de este documento usted debe ingresar a la
     siguiente url https://validador.sea.gob.cl/validar/2162850911" o
    similares está al final de cada página y no debe considerarse.
- Te proporcionaré algunos ejemplos de entradas y salidas esperadas. Úsalos
    como guía para procesar nuevas entradas.
- NO confundas el "numero de tabla" con el número de referencia del ICE o
    otros capítulos referenciados en la tabla. Si no estás seguro de haberlo
    identificado, guardalo cómo "no identificado".
```

c) **Formato de Salida**: se especifica el formato de respuesta *Pydantic* que deberá entregar el modelo, el cuál será diferente para cada considerando, en función de los atributos presentes en cada tabla. Para reforzar y aclarar al modelo que debe ir en cada variable, se entregan descripciones de cada campo.

d) **Ejemplos**: se entrega una serie de ejemplos para orientar al modelo, dependiendo de cada considerando, los cuáles muestran la entrada de ejemplo y la salida deseada para dicho caso.

7. **Formateo y Deduplicación de Datos:** El resultado del punto anterior es un conjunto de filas que representan cada considerando, con potenciales duplicados debido al solapamiento de *chunks*. Sobre esta, se aplica nuevamente *gpt-4o-mini* para eliminar duplicados y consolidar información, siguiendo ejemplos e instrucciones de deduplicación:

```
## Contexto:
Te proporcionaré un archivo JSON que contiene un conjunto de filas, cada una
    describiendo obligaciones presentes en una Resolución de Calificación
    Ambiental (RCA) de un proyecto de inversión en Chile.
```

```
## Objetivo:
Tu tarea es eliminar las filas duplicadas, es decir, las filas con el mismo valor
    en la variable "numero_tabla".
- Si hay diferencias en los valores de estas filas duplicadas, deberás combinarlas
     para formar una fila completa y coherente. El JSON resultante deberá cumplir
    con los formatos especificados para cada variable, y cada fila deberá estar
    ordenada de acuerdo con la estructura de la clase `FINAL`.
- El resultado final siempre deberá tener una sola fila por cada "numero_tabla".
- Debes preservar la máxima fidelidad al texto original. Si una de las filas
    duplicadas contiene parte del contenido y la otra tiene otra porción, debes
    combinarlas sin repetir el contenido.
- Si una variable está vacía en una fila duplicada, utiliza el contenido
    disponible en la otra fila. Si ambas están vacías, mantén el campo vacío.
- La variable "seccion" debe mantenerse exactamente como está en las filas
    originales.
- Las filas que no tienen duplicados deben mantenerse inalteradas.
Te proporcionaré algunos ejemplos de entradas y salidas esperadas. Úsalos como
    guía para procesar nuevas entradas.
```

8. **Almacenamiento de Resultados:** Los datos procesados se guardan en el archivo intermedio.

### Paso 2: Extracción de Obligaciones de los Considerandos

- **Clasificación de Considerandos:** Se dividen los considerandos en dos grupos para un tratamiento específico:
  - **Grupo Único:** 'PAS' y 'Contingencias y Emergencias', donde se asume una obligación por cada fila.
  - **Grupo Múltiple:** Resto de los considerandos, donde puede haber múltiples obligaciones por cada fila.
- **Aplicación del Modelo GPT:** Se utiliza el modelo *gpt-4o-mini*, con instrucciones y ejemplos diferentes para cada grupo, por ejemplo, para el Grupo Único:

```
Objetivo:
       El objetivo es categorizar las filas
presentes en el archivo de entrada organizadas bajo diferentes 'numeros_tabla'.
       Por cada fila de entrada debe generarse
una fila de salida, con las siguientes dimensiones:
...
(otros detalles específicos)
```

```
...

       Detalles Finales
       - Si hay variables de entrada que no
tienen texto, es irrelevante, debes cumplir tu objetivo igual.
       - Si alguna categoría no es clara,
elige la opción más cercana temática y semánticamente.
       - Recuerda, siempre usa una fila de
entrada por una fila de salida, reflejando con precisión la intención y el
contexto del texto original.
```

- **Formateo y Deduplicación de Datos:** Nuevamente, se utiliza *gpt-4o-mini* para eliminar duplicados semánticos y asegurar consistencia, entre las obligaciones que pertenezcan a un mismo subconsiderando. Para ello se entregan instrucciones como la siguiente:

```
## Contexto:
   Te proporcionaré un archivo JSON que
contiene un conjunto de filas, cada una describiendo obligaciones presentes en
una Resolución de Calificación Ambiental (RCA) de un proyecto de inversión en
Chile.

   ## Objetivo 1:
   Tu tarea es eliminar las filas duplicadas,
es decir, las filas que contienen una misma obligación.
   - Es posible que existan diferencias
semánticas en la manera en la que una misma obligación es expresada. Si ello
ocurre, deberás combinar todas las filas que tengan la misma obligación (pueden
ser más de 2 incluso) para formar una fila completa y coherente.
   - El JSON resultante deberá cumplir con los
formatos especificados para cada variable, y cada fila deberá estar ordenada de
acuerdo con la estructura determinada.
   - El resultado final siempre deberá contar
con un número igual o menor de filas que el recibido. Nunca se deberá entregar
una tabla vacía como resultado.
   - Debes preservar la máxima fidelidad al
texto original. Si una de las filas duplicadas contiene parte del contenido y
la otra tiene otra porción, debes combinarlas sin repetir el contenido.
```

```
   - Si No estás seguro de si una fila es un
duplicado, es preferible eliminarla y combinarla con otra.

   ## Objetivo 2:
   Asegurar consistencia:
   - Las variables de
"componente_materia", "fuente", "numero_fuente"
SIEMPRE deben tener el mismo valor para todas las filas entregadas, si una fila
no cumple, debes forzar para que tenga lo mismo que la mayoría de las filas.
   - Si hay alguna obligación
"Secundaria" entonces debe existir alguna obligación
"Principal", según lo que estimes más adecuado.
   Te proporcionaré algunos ejemplos de
entradas y salidas esperadas. Úsalos como guía para procesar nuevas entradas.

   Debes evitar a toda costa tener
obligaciones duplicadas, esa es tu prioridad!
```

- **Almacenamiento de Resultados:** El modelo genera un listado de obligaciones que se guardan en el archivo de obligaciones, replicando variables originales e incorporando nuevas categorías según las instrucciones.

## Metodología Alternativa:

### Paso 1: Extracción de Subconsiderandos de los Archivos

El primer paso del procedimiento consiste en transformar los archivos PDF de las Resoluciones de Calificación Ambiental (RCA) en insumos procesables de manera sistemática. Para ello, se emplea el servicio *Amazon Textract*, que permite digitalizar de forma estructurada los documentos. El sistema extrae no sólo texto, sino también tablas, encabezados, pies de página y otros elementos de formato. Cada fragmento se almacena como un bloque único con un identificador, lo que posibilita una manipulación flexible en etapas posteriores.

- **Segmentación de Considerandos:** Una vez extraído el contenido, se procede a identificar los distintos considerandos normativos. Para ello, se construyó un diccionario de encabezados frecuentes en este tipo de documentos (los mismos utilizados en la metodología anterior). Sobre esta base, se aplicó un algoritmo de coincidencia difusa (*fuzzy matching*), que compara los párrafos reconocidos con los encabezados esperados, tolerando variaciones de formato o redacción. De esta manera, se delimita cada considerando y se descartan los bloques irrelevantes, tales como pies de página o encabezados repetidos.

- **Segmentación de texto y tablas:** El siguiente paso consiste en separar los bloques de texto libre de las tablas incluidas en los documentos.
    - En el caso de las tablas, se identifican y, cuando éstas se encuentran divididas en varias páginas, se aplican rutinas de unión automática para reconstruirlas en un único objeto coherente. Posteriormente, cada tabla es convertida a un formato estructurado (JSON), que conserva su numeración (e.g., "Tabla 3.2") utilizando expresiones regulares (*regex*).
    - En el caso de los párrafos en texto libre, cada bloque se almacena de manera independiente como un subconsiderando, lo que asegura granularidad para su análisis posterior.
- **Identificación de Considerandos Relevantes.** Con el objetivo de enfocar el análisis en las obligaciones de carácter ambiental, se definió un conjunto de considerandos normativamente pertinentes, tales como:
    1. Medidas de Mitigación
    2. Permisos Ambientales Sectoriales
    3. Forma de Cumplimiento
    4. Condiciones y Exigencias
    5. Compromisos Voluntarios
    6. Contingencias y Emergencias
    7. Plan de Seguimiento

    Cada considerando se localiza dentro de la RCA a partir de los encabezados identificados en el paso de coincidencia difusa. De este modo, se establecen los límites iniciales y finales de cada considerando, permitiendo separar los bloques de texto y las tablas que le corresponden.
- **Almacenamiento de Bloques de Resultados:** Al finalizar este paso, cada RCA queda representada como un conjunto de subconsiderandos normalizados: fragmentos de texto y tablas estructurados, clasificados en los considerandos normativas relevantes. Esta representación constituye la base para las etapas siguientes de estandarización y categorización de obligaciones.

### Paso 2: Extracción de Obligaciones de los Considerandos

En esta etapa, se diseñó un procedimiento automatizado para identificar y extraer las obligaciones contenidas en los considerandos de las resoluciones ambientales. El procedimiento utiliza *gpt-4o-mini*, tomando como entrada un bloque de texto de la resolución y generando como salida un JSON válido que cumpliese con un esquema previamente definido. La siguiente instrucción fue utilizada como base para la categorización, la cuál se combina con instrucciones especificas para cada considerando.

```
Tarea: Se le entregará un texto de un documento de revisión
ambiental para un proyecto de inversión. Su objetivo es extraer
todos los requisitos de cumplimiento contenidos en el texto.
Haga coincidir el texto devuelto del requisito lo más cerca posible
del texto de entrada, pero sin perder la concisión.
Puede haber cero o más requisitos de cumplimiento.

Para cada requisito de cumplimiento identificado,
también identificará las siguientes características:

1. Componente ambiental: Componente del medio ambiente que el requisito
de cumplimiento busca proteger. Puede ser uno de los siguientes valores:
...

2. Fase Proyecto: Fase del proyecto a la que pertenece el requisito de cumplimiento.
Puede ser uno o más de los siguientes valores:
...
(continua con otras variables)
...
Para cada requisito que identifiques, debes clasificarlo según el siguiente esquema:
```

Cada bloque de texto es procesado y, en caso de contener obligaciones, se genera una fila en la base de datos con la información estructurada. El resultado final es un conjunto de obligaciones estandarizadas, cada una asociada a su numeral y considerando de origen en la resolución.

A diferencia de la Metodología Principal, acá no es necesario la eliminación de duplicados, puesto que cada uno de los bloques de información relevante identificados mediante *textract*, se procesa una sola vez.

## Post-Procesamiento de ambas metodologías

Posteriormente se realizan ajustes adicionales para garantizar la coherencia y completitud de los datos en todas las filas. Es importante destacar que el código incluye manejo de errores mediante bloques *try/except*, lo que permite reintentar la ejecución de funciones ante fallos o implementar alternativas cuando sea necesario. No obstante, no es posible asegurar que el código esté exento de errores, particularmente con los archivos que son imágenes escaneadas.

a. **Nº PAS:** se restringe el rango de número de PAS identificado para que esté entre el 111 y el 161, de manera que coincidan con su numeración en el reglamento del SEIA. Para los casos fuera de ese rango, mediante el modelo

de *embeddings text-embedding-3-small* se compara el contenido de cada registro con el contenido de cada uno de los permisos ambientales sectoriales, y si se detecta una similitud alta, entonces se modifica el número de PAS al correspondiente.

b. **Condiciones PAS:** mediante el modelo de *embeddings text-embedding-3-small* se comparan semánticamente las condiciones adicionales de cada uno de los PAS que indican que sí establecen condiciones, con el decreto respectivo, y se define, sobre cierto umbral, si las condiciones son equivalentes o no a las establecidas por la normativa vigente.

c. **Contingencias y Emergencias:** mediante el LLM *gpt-4o-mini* se clasifica cada obligación del considerando de contingencias y emergencias en una tipología y una sub-tipología de riesgos, tales como "Riesgos Operacionales", "Riesgos Naturales" y "Riesgos Antrópicos".

d. **Fuentes Normativas:** se estandariza y unifica el formato de las fuentes normativas de cada obligación, unificando textos como "19.300/1994", "1930094", "19300", "LEY 19300 / 1994" en un solo elemento, eliminando diferencias y errores de tipeo.

## 2. Anexo 2: Validación de los Resultados Obtenidos por IA

El presente anexo describe el mecanismo de validación utilizado para evaluar el desempeño del modelo de IA en la tarea de identificación de obligaciones contenidas en Resolución de Calificación Ambientals (RCAs).

La metodología de validación implementada se alinea con las prácticas y hallazgos reportados en la literatura de LLMs y procesamiento de lenguaje natural (NLP) aplicados en dominios como el legal y el clínico. Un elemento central en la literatura, y que también se aplica en este estudio, es el uso de un *ground truth* o *gold standard* manual y la revisión experta para evaluar los resultados del modelo. Además, se emplean métricas estándar de desempeño: *Recall* y *Precision*.[9]

Otro elemento fundamental, es que la validación se realiza bajo un enfoque *out of sample*, es decir utilizando datos no empleados durante la fase de entrenamiento del modelo. En particular, la IA fue entrenada exclusivamente con RCAs del sector minero, mientras que la validación se lleva a cabo con RCAs del sector energético. Esta separación entre los datos de entrenamiento y de evaluación permite mitigar el riesgo de *overfitting*, y proporciona evidencia de que el modelo es capaz de generalizar adecuadamente frente a documentos no vistos y con estructuras normativas distintas.

Además, la muestra para la validación se construye usando un muestreo aleatorio estratificado, considerando criterios como el tipo y formato del documento, reforzando la representatividad y validez estadística de los hallazgos.

## Modelo Estadístico

Considerando una serie de documentos (RCAs), cada uno de los cuales contiene un conjunto de obligaciones relevantes, distribuidas en distintas secciones. La tarea del modelo es clasificar estas obligaciones usando un modelo de clasificación no supervisado. Para evaluar la calidad de la clasificación, se utiliza un conjunto de documentos de prueba que no fueron usados para entrenar el modelo. A este conjunto se le asignan etiquetas manuales mediante revisión experta, generando un *ground truth* que sirve como referencia para la validación.

Para medir la calidad de las predicciones del modelo, se construyen muestras aleatorias de la población total y se evalúan dos métricas estándar:

- **Recall:** Proporción de elementos correctamente identificados relativo al total de elementos presentes en el texto. Indica la capacidad del modelo para identificar todas las obligaciones existentes en las RCAs.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

[9] (1), (3), (4), (5).

Donde,

- *True Positives* (TP): obligaciones presentes en el *ground truth* que fueron correctamente identificadas por la IA.
- *False Negatives* (FN): obligaciones presentes en el *ground truth* que no fueron reconocidas por la IA.

- **Precision:** Proporción de elementos correctamente identificados, relativo al total de elementos identificados por el modelo. Mide la fiabilidad de las predicciones positivas del modelo.

$$\textit{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Donde,

- *True Positives* (TP): obligaciones correctamente identificadas por la IA y efectivamente presentes en el documento.
- *False Positives* (FP): elementos marcados como obligaciones por la IA que no corresponden a obligaciones reales según el criterio experto.

Para el cálculo de estas métricas se utiliza el siguiente enfoque:

- Para el *recall*, se toma una muestra aleatoria de obligaciones presentes en las RCAs, sin intervención del modelo, y se evalúa si fueron correctamente identificarlas por la IA.
- Para la *precision*, se toma una muestra aleatoria de obligaciones extraídas por la IA y se verifica manualmente si corresponden efectivamente a obligaciones reales contenidas en el texto de las RCAs.

Ambas muestras se construyen mediante un muestreo aleatorio estratificado, seleccionando primero documentos (RCAs) según criterios relevantes, y luego extrayendo obligaciones dentro de cada documento. Como tanto la *precision* como el *recall* son proporciones entre 0 y 1, el error muestral se estima bajo un modelo binomial, lo que permite construir intervalos de confianza estadísticamente válidos para cada métrica.

## Tamaño Muestral y Criterios de Estratificación

Para ambos componentes de validación, *recall* y *precision*, se aplican criterios estadísticos estándar para determinar el tamaño muestral mínimo requerido. Se utiliza el siguiente cálculo estadístico, basado en un intervalo de confianza de una distribución binomial:

$$n = \frac{Z^2 \cdot p \cdot (1 - p)}{e^2}$$

**Cuadro 3: Cálculo de Tamaño Muestral Mínimo**

| Parámetro | Valor |
|---|---|
| Nivel de confianza | 95 % ($Z = 1,96$) |
| Error ($e$) | 0,05 |
| Proporción esperada ($p$) | 0,5 |
| Tamaño muestral estimado ($n$) | 384 |

Fuente: Elaboración Propia

Este valor representa el número mínimo de observaciones necesario para estimar una proporción con un nivel de confianza de 95 % (usando $Z = 1{,}96$), un margen de error $e$ con una proporción esperada $p$. En el Cuadro 3 se evalúa $n$ para $e = 5\,\%$ y $p = 0{,}5$ (el valor mas conservador). En la práctica, se supera este umbral tanto para el cálculo de *recall* como de *precision*, utilizando muestras de más de 384 obligaciones en ambos casos, lo que garantiza una estimación estadísticamente robusta de estas métricas.

El muestreo aleatorio estratificado, para las RCAs, se realiza considerando dos criterios principales:

- **Tipo de evaluación ambiental:** Si son Estudio de Impacto Ambientals (EIAs) o Declaración de Impacto Ambientals (DIAs), ya que afectan la estructura de las RCAs y la cantidad de secciones contenidas.
- **Formato del documento:** Si son escaneados o digitales, dado que para documentos escaneados se aplica un paso adicional de OCR.

La muestra utilizada para la validación se compone de 73 RCAs del sector energético, seleccionadas mediante muestreo aleatorio estratificado, como se detalla en el Cuadro 4:

**Cuadro 4: Composición de la Muestra de RCAs**

| Categoría | Digital (no escaneada) | Escaneada |
|---|---|---|
| Declaración de Impacto Ambiental (DIA) | 43 | 14 |
| Estudio de Impacto Ambiental (EIA) | 10 | 6 |
| Total | 53 | 20 |

Fuente: Elaboración Propia

Este diseño busca mejorar la representatividad de la muestra, además de reducir la varianza, dado que las RCAs no presentan una estructura uniforme y las obligaciones no se distribuyen homogéneamente entre secciones. Según la experiencia de los expertos y la observación directa de las RCAs, ciertas secciones concentran una mayor cantidad de obligaciones, como forma de cumplimiento y compromisos voluntarios, mientras que otras, como mitigación, plan de seguimiento o contingencias o emergencias, presenta una densidad menor. Además, las RCAs que evalúan EIAs tienden a tener más secciones de interés (hasta 7), en comparación con las DIAs, que incluyen hasta cinco secciones de interés.

En base a esta distribución heterogénea de la densidad normativa observada en las distintas secciones de las RCAs, se establece un criterio fijo para la cantidad de obligaciones a extraer por sección. Este criterio fue definido a partir de la experiencia con RCAs y tiene como objetivo asegurar una muestra equilibrada y representativa (Cuadro 5).

**Cuadro 5: Criterio de Extracción de Obligaciones por Sección**

| Sección | Obligaciones extraídas |
|---|---|
| Forma de cumplimiento | 3 |
| Compromisos voluntarios | 2 |
| Condiciones o exigencias | 1 |
| Contingencias o emergencias | 1 |
| Medidas de mitigación | 1 |
| Plan de seguimiento | 1 |
| PAS | 1 |

Fuente: Elaboración Propia

Este criterio se aplica a cada RCA seleccionada, considerando únicamente las secciones efectivamente presentes en cada documento. Dado que las DIAs pueden contener hasta cinco de estas secciones, y los EIAs hasta siete, el número total de obligaciones extraídas varía según el tipo de RCA y su contenido efectivo. Esta metodología permite reflejar de manera más precisa la distribución real de obligaciones en el universo de RCAs.

Dado que no todas las RCAs contienen todas las secciones de interés, el número final de obligaciones extraídas no corresponde exactamente al total teórico esperado según el diseño de extracción por sección. Sin embargo, la muestra final supera el umbral estadístico mínimo de 384 observaciones, y presenta una distribución que refleja adecuadamente la diversidad de casos del universo de RCAs del sector energético, tanto en términos de tipo de evaluación como de formato documental.

## Validación del recall

La validación del *recall* busca medir cuántas de las obligaciones realmente presentes en las RCAs son correctamente identificadas por el modelo de IA. Es decir, estima la cobertura del modelo respecto del universo de obligaciones existentes en los documentos.

Para construir esta estimación, se utiliza la muestra de 73 RCAs del sector energético descrita anteriormente. Desde esta muestra, se extraen manualmente obligaciones por parte de expertos humanos, conforme al criterio de selección previamente definido y aplicando un procedimiento de aleatorización en la elección de las obligaciones. La extracción se enfoca exclusivamente en las secciones de interés, y el número de obligaciones por sección es ajustado según su densidad normativa.

La identificación de obligaciones se realiza por dos expertos independientes, quienes aplican criterios jurídicos e inter-

pretativos para determinar qué fragmentos del texto corresponden efectivamente a obligaciones fiscalizables. Debido a que las RCAs no presentan un punteo estructurado ni estandarizado, esta tarea exige un ejercicio detallado de lectura y juicio experto. En caso de discrepancias, se recurre a un tercer revisor para consolidar una versión definitiva del *ground truth*.

Una vez construida la base manual, se compara con las obligaciones detectadas por la IA en los mismos documentos. El *ground truth* final está compuesto por un total de 554 obligaciones, extraídas manualmente desde las 73 RCAs seleccionadas, superando el umbral muestral mínimo requerido (n = 384).

Posteriormente, cada experto evalua de forma independiente si la IA identifica correctamente cada una de las obligaciones del *ground truth*. Los resultados individuales fueron los siguientes:

- **Experto 1:** TP = 501, FN = 53, *Recall* = 90,4 %
- **Experto 2:** TP = 530, FN = 24, *Recall* = 95,7 %

Para el cálculo final se adopta un enfoque conservador: si al menos uno de los expertos considera que una obligación no había sido reconocida por la IA (FN), se mantiene dicha clasificación. Aplicando este criterio, se obtiene un total de TP = 500 y FN = 54, lo que resulta en un *recall* de 90,3 %. Se construye además un intervalo de confianza del 95 % para esta estimación, obteniendo un rango de [0,878 ; 0,927].

El valor de *recall* obtenido por el modelo (0,902) se sitúa dentro de los rangos superiores reportados en la literatura para tareas de extracción desde textos no estructurados. En el ámbito normativo, Zhang y El-Gohary (2016, 2017), reportan desempeños de *recall* de 0,944 y 0,987, respectivamente, utilizando técnicas de procesamiento de lenguaje natural (NLP). Martin et al. (2024) muestran un rango de 0,823 a 0,917, dependiendo del modelo de lenguaje de gran escala (LLM) utilizado y Li et al. (2024), presenta valores entre 0,556 a 1 según el tipo de información clínica extraída. Wornow et al. (2024), también en el ámbito de la salud, reportan un *recall* de 0,92 al aplicar LLMs para tareas complejas de clasificación textual.

En este contexto, el resultado de nuestro modelo se considera consistente con estándares internacionales, evidenciando una alta capacidad para recuperar obligaciones efectivamente presentes en documentos regulatorios como las RCAs.

## Validación de precision

La validación de *precision* busca estimar cuántas de las obligaciones que el modelo de IA identifica como válidas corresponden, efectivamente, a obligaciones reales contenidas en los textos de las RCAs. Es decir, evalúa la fiabilidad de los resultados positivos reportados por el modelo.

Para esta validación, se selecciona una muestra aleatoria de obligaciones entre aquellas extraídas por la IA, siguiendo

criterios de proporcionalidad según tipo de RCA (DIA/EIA) y número de obligaciones por sección. La muestra final se compone de 564 obligaciones, lo que supera el tamaño muestral mínimo requerido (n = 384) para estimaciones con un nivel de confianza del 95 % y un margen de error del 5 %.

Cada una de estas obligaciones es evaluada de forma independiente por dos expertos, quienes determinan si efectivamente la obligación extraída está presente en el texto de la RCA correspondiente y cumple con los criterios de fiscalización definidos.

Los resultados individuales obtenidos son los siguientes:

- **Experto 1:** TP = 552, FP = 12, *Precision* = 97,8 %
- **Experto 2:** TP = 544, FP = 20, *Precision* = 96,5 %

Para consolidar los resultados se utiliza un enfoque conservador: si al menos uno de los dos expertos considera que una obligación era un falso positivo, se clasifica como tal. Aplicando este criterio, se obtiene un total de TP = 539 y FP = 25, resultando en una *precision* final de 95,5 %.

Se construye un intervalo de confianza del 95 % para esta estimación, obteniendo un rango de [0,939 ; 0,973]. Este resultado indica que más del 95 % de las obligaciones extraídas por la IA fueron correctamente identificadas.

La *precision* obtenida por el modelo se encuentra dentro del rango superior observado en estudios comparables. En el ámbito normativo, Zhang y El-Gohary (2016) reportan un valor de 0,969 y en su estudio posterior (2017) un valor de 0,876. En contextos clínicos, Li et al. (2024) presentan un rango de *precision* entre 0,819 y 0,969, y Wornow et al. (2024) reportan un valor de 0,91. Martin et al. (2024), al evaluar distintos modelos de lenguaje, observan una mayor variabilidad, con valores entre 0,531 y 0,958. En este marco, la *precision* alcanzada por el modelo se ubica en la parte alta de los resultados reportados, lo que sugiere un buen nivel de exactitud en la identificación de obligaciones desde las RCAs.

## Validación de la clasificación

Una vez validada la extracción de las obligaciones, se evalua la calidad de la clasificación realizada por el modelo de IA sobre esas mismas obligaciones. Esta validación se limita, exclusivamente, al subconjunto de obligaciones correctamente identificadas por el modelo en el proceso de *recall*, lo que garantiza que los resultados reflejan únicamente la capacidad clasificatoria de la IA, sin interferencias derivadas de errores de extracción.

Al igual que en etapas anteriores, se construye un *ground truth* específico para esta tarea, correspondiente a la clasificación de las obligaciones ya validadas. Esta labor es realizada por dos expertos independientes. En los casos en existe discrepancia entre ellos, se recurre a un tercero para resolver las diferencias.

Se evalúan un total de siete variables de clasificación, agrupadas en cuatro tipos metodológicos según su estructura lógica: variables binarias, variables categóricas de selección única, variables multietiqueta y variables abiertas. Las

variables evaluadas son: componente o materia, orden, tipo, establece condiciones, fase, fuente y número de fuente. A cada tipo metodológico se le aplicó una estrategia de validación específica, y se estimaron los niveles de coincidencia entre el sistema automatizado y la codificación realizada por expertos humanos.

Para la consolidación de resultados, se adopta un enfoque conservador. Cada obligación es clasificada de manera independiente por dos expertos, lo que genera dos evaluaciones paralelas. Solo en los casos en que ambos expertos coinciden en que la clasificación de la IA era correcta, se considera válida. En cambio, cuando existe discrepancia -es decir, si al menos uno de los expertos evalúa la clasificación como incorrecta-, esta se registra como error. Este criterio busca asegurar la máxima rigurosidad en la estimación, evitando sobrestimar el nivel de coincidencia del modelo.

## Variables binarias

Las variables binarias son aquellas en las que cada obligación puede clasificarse exclusivamente en una de dos categorías mutuamente excluyentes. En este análisis se consideran tres variables de este tipo:

- **Orden:** obligación principal o secundaria.
- **Tipo:** obligación de medios u obligación de resultados.
- **Establece condiciones:** indica si la obligación impone condiciones en el contexto de una Permiso Ambiental Sectorial (PAS).

Las variables de orden y tipo son evaluadas en obligaciones pertenecientes a las secciones de: forma de cumplimiento, compromisos voluntarios, contingencias o emergencias, medidas de mitigación y plan de seguimiento. En cambio, la variable establece condiciones se aplica exclusivamente a las obligaciones provenientes de la sección PAS, por ser el único contexto en que dicha distinción es relevante.

La validación consiste en comparar directamente la categoría asignada por la IA con la contenida en el *ground truth*. Técnicamente, esta comparación se expresa como *accuracy*, es decir, la proporción de clasificaciones correctas sobre el total de decisiones realizadas por el modelo. Los resultados bajo el enfoque conservador son:

- **Orden:** 82.2 % de *accuracy* (n=377). Con un intervalo de confianza de [0,784 ; 0,873].
- **Tipo:** 83.5 % de *accuracy* (n=376). Con un intervalo de confianza de [0,798 ; 0,873].
- **Establece condiciones:** 88.5 % de *accuracy* (n=61). Con un intervalo de confianza de [0,805 ; 0,965].

## Variable categórica de selección única

En esta categoría se valida la clasificación de la variable componente ambiental (también denominada materia ambiental), la cual representa el ámbito temático al que se vincula cada obligación. Esta variable contempla múltiples

categorías posibles (como aire, agua, emisiones, energía, flora y fauna, suelo, residuos, entre otras), pero cada obligación debe clasificarse únicamente bajo una de ellas.

Se aplican dos indicadores para su validación:

- **Accuracy:** o coincidencia exacta corresponde al porcentaje de veces en que la IA y el experto coinciden en la misma categoría. Es decir, la proporción de clasificaciones correctas sobre el total de obligaciones evaluadas.
- **Weighted accuracy:** o coincidencia ponderada, es la métrica que considera la proporción relativa de cada categoría en el conjunto total, ponderando los aciertos en función de su prevalencia. Este indicador es especialmente útil en contextos con muchas clases y distribución asimétrica.

Los resultados obtenidos bajo el criterio conservador son:

- **Accuracy:** 83.5 % (n=370). El intervalo de confianza es de [0,797 ; 0,873].
- **Weighted accuracy:** 95 %. El intervalo de confianza es de [0,927 ; 0,972].

Adicionalmente, se calcula el indicador de *accuracy* para las categorías de mayor prevalencia en el *ground truth* -flora y fauna y residuos y sustancias peligrosas- y la validación se realiza de manera independiente por dos expertos, optándose por reportar el perfil más conservador. Para flora y fauna se obtiene un 94,8 % de *accuracy* (intervalo de confianza: [0,929 ; 0,969]; n=400), categoría que representa el 32,2 % de las obligaciones en la muestra de este experto. En residuos y sustancias peligrosas se alcanza un 90,3 % de *accuracy* (intervalo de confianza: [0,873 ; 0,932]; n=400), con una prevalencia del 31,1 %.

## Variables multietiqueta

Dentro del conjunto de variables extraídas por el modelo de inteligencia artificial, dos de ellas, fase del proyecto y fuente jurídica, presentan una naturaleza multietiqueta. Esto significa que una misma obligación puede estar asociada a más de una categoría simultáneamente. Por ejemplo, una obligación puede aplicarse tanto a la fase de construcción como a la de operación, o bien sustentarse en múltiples fuentes jurídicas, como leyes, decretos y resoluciones administrativas. En el caso particular de las fuentes, se considera que cada obligación puede estar vinculada a un máximo de cuatro referencias normativas que constituyen la base para el cálculo de las métricas.

Esta complejidad impone desafíos particulares a la validación, dado que no basta con evaluar si una etiqueta está presente o no, sino que es necesario considerar el conjunto completo de etiquetas asignadas. Para abordar este problema se aplican métricas estándar en clasificación multietiqueta: *recall* y *precision*. De este modo, se reconoce que la IA puede identificar parcialmente las categorías correctas, aun cuando no logre replicar el conjunto completo definido en el *ground truth*. Por ejemplo, si el *ground truth* corresponde a {decreto, ley} y la IA predice solo {decreto}, se contabiliza

un acierto en *recall* (encuentra parte de las etiquetas), aunque la predicción no sea exacta. En contraste, si la IA asigna {decreto, ley, resolución}, obtiene un mayor *recall* pero penaliza en *precision* por incorporar una etiqueta adicional no presente en el *ground truth*. Este enfoque permite reflejar de manera más fiel los aciertos parciales y los errores de sobreclasificación.

Para el análisis se trabaja bajo el supuesto de independencia entre las etiquetas asignadas a una misma obligación, tanto en la variable fuente jurídica como en fase del proyecto. Esto significa que cada referencia normativa o fase fue considerada de manera independiente y sin correlación con las demás, lo que permite evaluar el desempeño del modelo como si cada etiqueta representara una unidad de clasificación autónoma.

Los resultados obtenidos por cada experto se resumen a continuación:

- **Experto 1**
  - Fuente jurídica: *Recall* = 91 % (n=405); *Precision* = 89 % (n=368).
  - Fase del proyecto: *Recall* = 94 % (n=725); *Precision* = 96 % (n=710).
- **Experto 2**
  - Fuente jurídica: *Recall* = 90 % (n=432); *Precision* = 83 % (n=471).
  - Fase del proyecto: *Recall* = 94 % (n=772); *Precision* = 94 % (n=770).

Adicionalmente, y siguiendo el mismo enfoque conservador aplicado anteriormente, se consolidan los resultados de ambos expertos considerando que, en caso de discrepancia, la clasificación de la IA se registra como incorrecta. Bajo este criterio, se obtienen los siguientes valores de *recall* y *precision*:

- **Fuente jurídica**
  - *Recall*: 90,6 % (n=405). Intervalo de confianza de [0,878 ; 0,935].
  - *Precision*: 89,1 % (n=367). Intervalo de confianza de [0,859 , 0,923].
- **Fase del proyecto**
  - *Recall*: 93,9 % (n=724). Intervalo de confianza de [0,922 ; 0,957].
  - *Precision*: 95,5 % (n=710). Intervalo de confianza de [0,940 ; 0,970].

## Número de fuente jurídica

Finalmente, se evalúa la capacidad del modelo para identificar correctamente el número específico de la norma jurídica, lo que aplica únicamente a aquellas obligaciones sustentadas en fuentes normativas explícitas (como leyes o decretos).

En cambio, categorías como compromisos voluntarios o decisiones de la autoridad no contemplan un número de fuente, por lo que quedan excluidas de este análisis. Es importante precisar que el *n* reportado en esta sección no corresponde al número de obligaciones, sino a la cantidad de fuentes normativas identificadas en cada caso, ya que cada obligación puede estar vinculada a una o más disposiciones legales.

La validación del número de fuente se realiza considerando únicamente los casos clasificados como verdaderos positivos en *recall* y en *precision*, de modo de asegurar que la comparación se restrinja a las instancias en que efectivamente corresponde identificar un número normativo.

Los resultados obtenidos son los siguientes:

- **Experto 1**
  - Número de fuente (*recall*): 90,3 % (n=195).
  - Número de fuente (*precision*): 84,5 % (n=181).
- **Experto 2**
  - Número de fuente (*recall*): 86,1 % (n=231).
  - Número de fuente (*precision*): 87 % (n=231).

De manera complementaria, se aplica un criterio conservador en la consolidación de resultados, siguiendo la misma metodología utilizada en etapas previas:

- Número de fuente (*recall*): 90,1 % (n=191). Intervalo de confianza [0,858 ; 0,943].
- Número de fuente (*precision*): 85,5 % (n=179). Con un intervalo de confianza de [0,803 ; 0,906].

A continuación, en el Cuadro 6, se presentan los resultados resumidos de la validación de las diferentes variables clasificatorias evaluadas, mostrando los indicadores de coincidencia obtenidos bajo el enfoque conservador y el tamaño de muestra asociado a cada análisis.

## Síntesis y conclusiones

La validación de la herramienta de IA desarrollada para la extracción automática de obligaciones desde RCAs se realiza siguiendo estándares metodológicos rigurosos, tanto en el diseño estadístico como en la verificación manual de resultados.

Se utiliza un enfoque *out of sample*, validando el modelo con RCAs del sector emergético. Esto permite evaluar la capacidad del modelo para generalizar a nuevos documentos y evitar posibles sesgos por *overfitting*. Sobre esta base,

**Cuadro 6: Resumen de Resultados de Validación de la Clasificación**

| Variable | Indicador | Resultado % | Muestra (n) | Intervalo de Confianza |
|---|---|---|---|---|
| Componente ambiental | *Accuracy*<br>*Weighted accuracy* | 83,5<br>95,0 | 370 | [0,797 ; 0,873]<br>[0,927 ; 0,972] |
| Tipo | *Accuracy* | 83,5 | 376 | [0,798 ; 0,873] |
| Orden | *Accuracy* | 82,2 | 377 | [0,784 ; 0,873] |
| Establece condición | *Accuracy* | 88,5 | 61 | [0,805 ; 0,965] |
| Fase | *Recall*<br>*Precision* | 93,9<br>95,5 | 724<br>710 | [0,922 ; 0,957]<br>[0,940 ; 0,970] |
| Fuente | *Recall*<br>*Precision* | 90,6<br>89,1 | 405<br>367 | [0,878 ; 0,935]<br>[0,859 ; 0,923] |
| Número de fuente | *Recall*<br>*Precision* | 90,1<br>85,5 | 191<br>179 | [0,858 ; 0,943]<br>[0,803 ; 0,906] |

Fuente: Elaboración Propia

se seleccionan muestras aleatorias estratificadas y se asegura un tamaño muestral suficiente para estimar métricas de desempeño con un nivel de confianza del 95 % y un margen de error aceptable. La revisión experta de las obligaciones permite construir un *ground truth* robusto, tanto para evaluar la capacidad del modelo de encontrar las obligaciones existentes (*recall*) como para medir la confiabilidad de las que declara haber encontrado (*precision*).

En ambos casos, los resultados superaron los requerimientos mínimos establecidos para validar el modelo. La evidencia empírica sugiere que el sistema de IA es una herramienta confiable para apoyar la caracterización de la carga regulatoria ambiental, y puede ser extendido a otros sectores o tipos de documentos con adecuaciones metodológicas equivalentes.

Adicionalmente, se evalúa la capacidad del modelo no solo para extraer obligaciones, sino también para clasificarlas en múltiples dimensiones relevantes. Se aplican criterios conservadores y procedimientos equivalentes de revisión experta, arrojando niveles altos de coincidencia entre la clasificación automática y la realizada por humanos. Esto refuerza la utilidad del modelo no solo como herramienta de extracción, sino también como insumo para tareas de análisis estructurado.

Más allá de los resultados específicos, este ejercicio pone en relieve la importancia de aplicar métodos de validación rigurosos al utilizar sistemas de IA en contextos regulatorios o de función pública. El uso de modelos de lenguaje se ha expandido rápidamente, pero en muchos casos sin procesos sistemáticos de verificación, lo que aumenta el riesgo de generar conclusiones erróneas o decisiones mal fundamentadas.

Este estudio contribuye a cerrar esa brecha, mostrando que es posible y necesario validar empíricamente el desempeño de sistemas automatizados antes de integrarlos en procesos institucionales.

Si bien las metodologías de validación aún están en desarrollo y varían según el tipo de aplicación o contexto, este estudio muestra que existen ciertos principios que pueden fortalecer significativamente la robustez y transparencia de los procesos de evaluación: validación *out of sample*, aleatorización, definición adecuada del tamaño muestral y uso de métricas. Estos elementos contribuyen a mejorar la confiabilidad técnica de los modelos y respaldan su legitimidad en contextos donde la transparencia y la rendición de cuentas son especialmente relevantes.

## Referencias


[1] Li, D., Kadav, A., Gao, A., Li, R., & Bourgon, R. (2024). Automated clinical data extraction with knowledge conditioned llms. *arXiv preprint arXiv:2406.18027*.

[2] Martin, L., Whitehouse, N., Yiu, S., Catterson, L., & Perera, R. (2024). Better call gpt, comparing large language models against lawyers. *arXiv preprint arXiv:2401.16212*.

[3] Wornow, M., Lozano, A., Dash, D., Jindal, J., Mahaffey, K. W., & Shah, N. H. (2025). Zero-shot clinical trial patient matching with llms. *NEJM AI, 2*(1), AIcs2400360.

[4] Zhang, J., & El-Gohary, N. M. (2016). Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking. *Journal of computing in civil engineering, 30*(2), 04015014.

[5] Zhang, J., & El-Gohary, N. M. (2017). Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking.*Automation in construction, 73*, 45-57.