

Fiscal HeredIA:

IA generativa al servicio de
la investigación penal



Richard Weber

Doctor en Investigación de Operaciones y Licenciado en Matemáticas por la Universidad de Aachen, Alemania. Profesor Titular del Departamento de Ingeniería Industrial de la Universidad de Chile. Lideró el desarrollo del sistema Fiscal HeredIA para el Ministerio Público, que fue galardonado con el premio Potencia Chile en enero de 2026 y actualmente continúa con el proyecto desde la Fiscalía de Chile.

✉ richard.weber@uchile.cl



Valentina Reyes

Ingeniera Civil Industrial por la Universidad de Chile, con doble título con la Escuela Central de Marsella, Francia. Analista criminal de la Unidad Coordinadora SAC de la División de Estudios de la Fiscalía Nacional, Ministerio Público de Chile. Se desempeñó como asistente de investigación en el marco del proyecto Fiscal HeredIA y actualmente continúa con el proyecto desde la Fiscalía de Chile.

✉ vreyes@minpublico.cl



Resumen / Desde 2023, el Ministerio Público chileno usa Fiscal Heredia, un ecosistema de inteligencia artificial que hoy emplean más de 700 fiscales y funcionarios del Ministerio Público. El sistema permite cargar documentos y audios de una investigación y formularles preguntas en lenguaje natural, con una restricción deliberada: las respuestas provienen únicamente de la evidencia cargada. En este artículo contamos cómo funcionan los módulos de análisis de texto y audio, qué problema operativo del oficio fiscal vinieron a resolver, y qué nos enseña sobre llevar la IA al sector público chileno.

Quince tomos

En una causa por fraude tributario, lavado de activos o crimen organizado, una carpeta investigativa puede llegar a los quince tomos. Cada tomo, mil doscientas a mil quinientas páginas. Hagamos la cuenta: una sola causa puede llegar a las veinte mil páginas de prueba documental. Declaraciones, peritajes, cartolas bancarias, capturas de redes sociales, comprobantes de transferencias, informes policiales, oficios.

Veinte mil páginas son, conservadoramente, cuarenta novelas. Imagine leerlas todas. Ahora imagine leerlas todas y además recordar quién dijo qué, en qué tomo, en qué página, para responder preguntas como: *¿se repiten cuentas receptoras entre los doscientos casos?*, *¿qué dijo el imputado en la escucha del 14 de marzo que contradice su declaración?*, *¿hay alguien que aparezca conectado con tres o más de los detenidos?*

Y eso es sólo el papel. A las veinte mil páginas se suman, en investigaciones de este tipo, escuchas telefónicas autorizadas judicialmente que pueden superar las quinientas horas de audio. Veinte días continuos, sin dormir, escuchando.

A la velocidad de la lectura humana, responder eso no toma semanas. Toma meses. Y el reloj corre: el artículo 247 del Código Procesal Penal otorga al fiscal dos años desde la formalización para cerrar la investigación. La jurisprudencia ha sido clara en que el “exceso de trabajo” no constituye fuerza mayor. La carga, simplemente, no es excusa.

Este artículo es sobre la herramienta que se construyó para intervenir exactamente ahí. Se llama Fiscal Heredia. Y lo más interesante no es que use inteligencia artificial, cualquiera puede hacer eso hoy, sino cómo la usa: con una restricción que la mayoría de las aplicaciones de IA evita, y que en este contexto resulta innegociable.

¿Qué es Fiscal Heredia?

El nombre no es casualidad. Heredia es el detective que Ramón Díaz Eterovic ha hecho caminar por las calles de Santiago durante más de tres décadas, resolviendo casos

a punta de caminatas, lecturas y una memoria atenta a los detalles que la ciudad va descartando. Cuando el Ministerio Público y la Universidad de Chile bautizaron así a su sistema de IA, eligieron una imagen: una herramienta que, como el detective, lee mucho, recuerda todo y conecta lo que parecía no estar conectado —y que también, casi de regalo, deja un acrónimo prolijo: HERramienta para el Estudio del Delito usando Inteligencia Artificial (HEREDIA).

Fiscal Heredia nació en 2020 como un proyecto FONDEF entre el Ministerio Público, la Universidad de Chile, la Universidad del Bío-Bío y la Universidad de los Andes. Tras un segundo FONDEF en 2024, hoy lo usan más de 700 fiscales, analistas y asesores en distintas regiones del país.

No es un modelo. Es un ecosistema modular: distintas piezas que resuelven problemas operativos distintos del oficio fiscal, montadas sobre una infraestructura común (ver Figura 1). En este artículo cubrimos dos: el análisis de texto y el de audios. Dejamos fuera los módulos de redes delictuales [1, 2] y videos.

Textos: preguntarle a una carpeta

Empecemos por lo más simple. Un fiscal tiene un archivo PDF —un informe pericial, un conjunto de comprobantes bancarios, una declaración extensa— y necesita extraer información. Carga el archivo, escribe una pregunta en lenguaje natural, recibe una respuesta. *“¿Qué cuentas bancarias se mencionan?”*, *“¿hay transferencias entre estas tres personas?”*, *“resume las declaraciones de la víctima X”*.

Detrás de esa interfaz simple ocurren dos cosas. Primero, el sistema reconoce el texto incluso cuando el PDF viene de imágenes como escaneos, fotos, capturas de pantalla. Sin ese paso, una transferencia fotografiada con el celular resulta invisible para el sistema. Segundo, un modelo de lenguaje grande (LLM por su sigla en inglés) interpreta la pregunta y elabora la respuesta a partir del texto extraído.

Acá conviene detenerse. Fiscal Heredia no opera como un buscador general ni como un asistente que responde desde Internet. Su diseño busca que las respuestas se apoyen



Figura 1 / Ecosistema Fiscal Heredia.

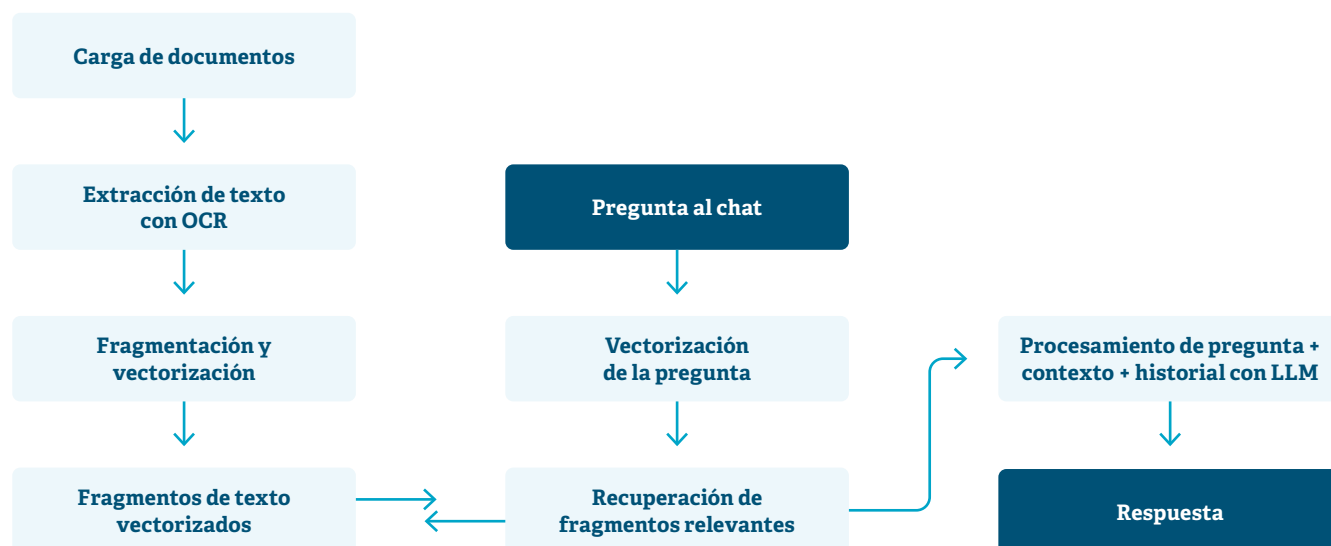


Figura 2 / Análisis de textos en Fiscal Heredia.

en los documentos incorporados al caso y que cada afirmación relevante pueda vincularse con antecedentes verificables. Esto no elimina el carácter probabilístico propio de los modelos de lenguaje, pero sí reduce el riesgo de que el sistema introduzca información ajena al corpus analizado. Para la investigación criminal, esa trazabilidad es central: los datos externos pueden ser útiles para orientar los análisis, formular hipótesis o sugerir nuevas diligencias, pero deben distinguirse claramente de la evidencia y de los antecedentes contenidos en la carpeta investigativa. La Figura 2 ilustra el funcionamiento del módulo de análisis de textos.

Cuando los documentos son muchos, el problema cambia de naturaleza. Quinientas carpetas no caben en la memoria de trabajo de ningún modelo de lenguaje actual. ¿Cómo se le pregunta a quinientas carpetas?

La idea es elegante en su descripción. Cada fragmento de texto se convierte en una representación numérica (*embedding*) que captura su significado, no sus palabras exactas. Esas representaciones se guardan en una base especializada. Cuando llega una pregunta, el sistema también la convierte en una representación numérica y busca los fragmentos cuyos *significados* se asemejan al de la consulta.



Figura 3 / Análisis de audios en Fiscal HeredIA.

Sólo esos fragmentos, los más relevantes, se le pasan al modelo para que componga la respuesta.

¿Qué es un *embedding*? Una representación numérica que ubica fragmentos de texto en un espacio donde la cercanía equivale a similitud de significado. Así, “*transferencia bancaria sospechosa*” y “*depósito atípico desde una cuenta no asociada*” terminan cerca, aunque no compartan ninguna palabra. La búsqueda deja de ser por coincidencia literal y pasa a ser por sentido.

Este patrón de recuperar primero los fragmentos pertinentes y generar después la respuesta tiene una virtud que importa especialmente acá. Cada respuesta queda anclada a fragmentos identificables de la evidencia. El o la fiscal puede pedir que el sistema le muestre exactamente de qué documentos y de qué páginas salió cada afirmación.

Audios: cuando la evidencia es una conversación

El cambio de medio plantea problemas distintos. Un audio no se lee, se escucha. Y rara vez está limpio: hay ruido, voces que se solapan, jerga, idiomas que no son los nativos. El módulo de análisis de audio lo aborda en tres movimientos [3].

Transcripción. Convertir habla en texto, robusto a ruido, con detección automática del idioma. Los idiomas soportados hoy son español, turco, chino y guaraní. Esa lista, por

sí sola, dice algo sobre el tipo de investigaciones donde el sistema entró en juego.

Diarización. Es el componente menos conocido fuera del rubro y, técnicamente, el más elegante. *Diarizar* es identificar quién habla en cada momento, sin saber quién es [4].

Traducción. Si el audio no está en español, se traduce automáticamente antes de continuar.

Una vez transcrito, diarizado y eventualmente traducido, el audio cambia de naturaleza: deja de ser un archivo que hay que escuchar de punta a punta y se convierte en material consultable. El o la fiscal puede buscar menciones, revisar intervenciones, identificar quién dijo qué o solicitar resúmenes de tramos extensos. La idea de ecosistema surge precisamente ahí: en la posibilidad de aplicar capacidades de análisis a distintos tipos de evidencia, desde documentos escritos hasta registros sonoros, sin reducir el trabajo investigativo a un solo formato. La Figura 3 muestra el funcionamiento del módulo de análisis de audios.

Lo que la herramienta no hace

¿Y si Fiscal HeredIA se equivoca?

Los modelos de lenguaje pueden producir respuestas plausibles que no estén suficientemente sustentadas en los documentos analizados. La restricción de fuente —responder a partir del contenido cargado— reduce ese riesgo, pero no lo

elimina. El sistema puede interpretar mal una pregunta, recuperar fragmentos poco pertinentes o formular una respuesta que parezca correcta sin estar adecuadamente respaldada por el documento. Por eso, la regla de uso es invariable: toda respuesta debe ser revisada y validada por una persona.

Esa validación no depende sólo de la etapa final. También importa cómo se usa la herramienta. Las respuestas son sensibles a la forma en que se pregunta: una consulta vaga tiende a producir una respuesta vaga; una pregunta precisa permite orientar mejor la búsqueda y revisar con más claridad las fuentes que sustentan el resultado. Por eso, la implementación de HeredIA no puede limitarse a otorgar acceso a una plataforma. Requiere capacitación, acompañamiento constante y espacios de retroalimentación con los usuarios, para que fiscales y analistas desarrollen criterios de uso, reconozcan los límites del sistema y aprendan a formular preguntas que sirvan realmente al trabajo investigativo.

Lo mismo ocurre con los audios. Las transcripciones automáticas no son actas. Ruido ambiental, voces solapadas, mala calidad de grabación o modismos locales pueden generar errores. El valor del sistema está en acelerar la revisión, ordenar el material y permitir búsquedas dirigidas; no en reemplazar la escucha crítica ni la comprobación humana cuando el contenido será usado en decisiones relevantes.

Por eso, la pregunta más importante cuando el Estado adopta herramientas de inteligencia artificial no es sólo qué tan buena es la tecnología. Es qué proceso humano, institucional y metodológico se construye alrededor de ella. Fiscal HeredIA debe entenderse en esa lógica: no como un sistema que impone respuestas, sino como una herramienta diseñada para acompañar el análisis, ampliar la capacidad de revisión y apoyar decisiones que siguen siendo responsabilidad de las personas.

Cuatro casos, cuatro módulos

Cifras de impacto hay muchas. Pero el valor de un sistema modular se entiende mejor al examinar casos en los que cada módulo entró en juego por separado.

Estafas masivas por Internet (2024). Doscientas carpetas de estafas cometidas con sitios web fraudulentos. Reducción del 97% del tiempo de análisis. El punto crítico: el ahorro de tiempo no es eficiencia administrativa. Es la diferencia entre que una formalización llegue oportunamente, dentro del plazo legal, o que la oportunidad se pierda.

Escuchas de audios (2024). Una investigación penal involucraba a una banda de extranjeros cuya lengua materna

Una sola causa puede llegar a las veinte mil páginas de prueba documental [...] Ahora imagine leerlas todas y además recordar quién dijo qué, en qué tomo, en qué página.

no era español. Durante meses, el material de las escuchas telefónicas autorizadas judicialmente permaneció prácticamente inaccesible: nadie en el equipo investigativo dominaba el idioma, y contratar traducción para cientos de horas de audio era inviable en los plazos disponibles. Con HeredIA, el módulo de transcripción y traducción procesó el material y lo convirtió en texto consultable en español.

Operación Influencers (2023). Una banda dedicada a encerronas, transmitidas por redes sociales. La fiscalía tenía dos imputados. Se le pidió al sistema que sugiriera a quiénes más buscar. Sugirió trece. Tras la verificación en terreno, once resultaron parte efectiva de la banda. Importante: la herramienta no respondió “*quiénes son los culpables*” —no es lo que hace—, sino “*estos son los nodos más conectados con los ya conocidos*”. El paso a una formalización siguió siendo trabajo humano.

Proyecto Prófugos (2024). A mitad de 2023 había aproximadamente 1.200 prófugos peligrosos. La metodología tradicional capturaba entre el 1% y el 3% de los buscados. Con apoyo del módulo de redes, la cifra subió al 40% dentro de un año, es decir más de cuatrocientas personas resultaron detenidas. Lo notable no es sólo el salto porcentual: es que una capacidad analítica que antes simplemente no existía pasó a estar disponible para una región completa.

Lo replicable

Fiscal HeredIA ha recibido reconocimientos públicos por su aporte a la modernización del trabajo investigativo. En 2025, el Ministerio de Hacienda lo destacó como una iniciativa relevante de inteligencia artificial aplicada a la transformación digital del Estado, en el marco del encuentro “Para la transformación del Estado: personas, confianza y futuro”. Luego, recibió el Premio Nacional de Inteligencia Artificial Chile Potencia 2026, organizado por el Ministerio de Ciencia, Tecnología, Conocimiento e Innovación junto a Fundación País Digital, en la categoría “Organizaciones”. Además, figura entre los seis finalistas del IFORS Prize for OR in Development 2026, que se presentará en la conferencia



de IFORS en Viena, en julio de 2026. Son hitos relevantes. Pero no son la conclusión interesante de este artículo.

La conclusión interesante es qué de esto es replicable y qué no. Las técnicas que usa HeredIA están disponibles para cualquiera. Reconocimiento de texto, modelos de lenguaje, embeddings, transcripción automática, diarización: todo está documentado, todo se puede licenciar o usar mediante APIs. No son, por sí solas, lo que hizo posible este sistema.

Lo que lo hizo posible fue una combinación específica que sí es replicable, pero exige tiempo: una alianza Universidad-Estado sostenida durante más de cinco años a través de dos proyectos FONDEF consecutivos; un equipo desarrollador estable dentro de la institución, capaz de entender tanto el oficio fiscal como las técnicas; casos de uso bien acotados antes de escalar; y una cultura de validación humana incorporada al flujo de trabajo desde el primer día.

Pero hay una pieza adicional, menos visible y probablemente más decisiva: escuchar la operación. HeredIA no nace de imaginar desde afuera cómo debería trabajar una fiscalía, sino de observar dónde se pierde tiempo, qué tareas se repiten, qué información cuesta encontrar, qué dolores enfrentan fiscales, analistas y equipos investigativos cuando deben revisar miles de páginas, audios, declaraciones o antecedentes dispersos. Esa escucha permite que la tecnología no llegue como una imposición, sino como una herramienta que acompaña prácticas existentes, las ordena y las potencia.

Esa es la lección útil para otros despliegues de IA en el sector público chileno. La parte difícil no es sólo construir el modelo ni conectar una API. Es sostener una relación de trabajo con los usuarios, traducir problemas operativos en funcionalidades concretas, capacitar de manera continua, recoger retroalimentación y ajustar la herramienta a medida que cambia la práctica. Cualquiera de esas piezas faltando, el sistema difícilmente llega a producción o se vuelve parte del trabajo cotidiano.

Referencias

- [1] Troncoso F, Weber R (2024) The Steiner tree Prosecutor: Revealing and disrupting criminal networks through a single suspect. PLoS ONE 19(12): e0312827.
- [2] Troncoso, F., Weber, R. (2025): Using AI to Identify Criminal Gang Members in Chile. IFORS Newsletter 20, No. 4, 9-12. <https://www.ifors.org/december-2025-issue/>.
- [3] Bain, M., Huh, J., Han, T., Zisserman, A. (2023): WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. Proc. Interspeech 2023, 4489-4493.
- [4] Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S. Narayanan, S. (2022): A Review of Speaker Diarization: Recent Advances with Deep Learning. *Computer Speech & Language* 72.
- [5] Fiscalía 2026: Cuenta Pública 2026 del Fiscal Nacional Ángel Valencia Vásquez. <https://www.youtube.com/watch?v=v2hUJRqmnQ>.

El ahorro de tiempo no es eficiencia administrativa. Es la diferencia entre que una formalización llegue dentro del plazo legal, o que la oportunidad se pierda.



Figura 4 / Premios que obtuvo Fiscal HeredIA, presentados por el Fiscal Nacional Ángel Valencia [5; minuto 44:20].

Esa es, finalmente, la invitación que deja Fiscal HeredIA. Hay mucha inteligencia artificial en laboratorios, papers, prototipos y pruebas de concepto; el desafío es llevarla adonde puede producir un impacto público concreto. Salir de la academia no significa abandonar el rigor, sino ponerlo a trabajar en problemas reales: expedientes imposibles de leer a tiempo, audios que nadie alcanza a revisar completos, información dispersa que debe ordenarse antes de que venza un plazo legal. Fiscal HeredIA muestra que esa transferencia es posible cuando la investigación se encuentra con la operación, cuando los equipos técnicos escuchan a quienes usan la herramienta y cuando el objetivo no es demostrar que una tecnología funciona, sino transformar una práctica institucional. **B**