DOCUMENTO DE CONVOCATORIA · 2026

BELMA

Benchmark Evaluativo Legal Mexicano de Inteligencia Artificial

Hacia un estándar abierto
para la evaluación de sistemas de IA legales en México.

RESUMEN

BELMA es una iniciativa abierta para construir el primer dataset legal mexicano anotado por la comunidad jurídica, con el propósito de evaluar de forma rigurosa, reproducible y pública el desempeño de los sistemas de inteligencia artificial aplicados al derecho. La iniciativa convoca a profesores, investigadores, abogados en ejercicio, equipos legales internos y estudiantes de los últimos semestres a participar en la definición metodológica, la curaduría del corpus y la anotación de las tareas que conformarán el benchmark.

El benchmark será gobernado por un Comité Técnico independiente con representación plural de academia, barras de abogados y la práctica jurídica mexicana. El dataset, la metodología, el código de evaluación y los resultados serán de acceso público. Cualquier persona, despacho, equipo de investigación o empresa podrá evaluar su sistema de inteligencia artificial contra el mismo estándar.

El presente documento describe el problema que motiva la iniciativa, los principios que la rigen, las fases de trabajo previstas y los mecanismos de participación. La metodología definitiva será resultado del trabajo del comité y se publicará para comentarios antes de cerrarse.

§ 1 · MOTIVACIÓN

El problema de medir la inteligencia artificial en el derecho mexicano.

En los últimos años han proliferado los sistemas de inteligencia artificial aplicados al derecho, tanto comerciales como de investigación, con propuestas que abarcan análisis de documentos, investigación jurídica, redacción de escritos y soporte a la práctica. Sin embargo, no existe en México un punto de referencia compartido que permita evaluar y comparar el desempeño de estos sistemas de manera rigurosa.

Cada proveedor presenta sus propias métricas. Cada despacho prueba a su manera. Los compradores carecen de un estándar para distinguir las herramientas que funcionan de aquellas que solamente comunican bien. La academia, por su parte, no dispone de una base común sobre la cual investigar el comportamiento de los modelos de lenguaje en tareas legales en español jurídico mexicano. La literatura comparada existe, pero está construida sobre ordenamientos extranjeros y rara vez es transferible al contexto local.

BELMA propone llenar ese vacío. La iniciativa busca consolidar un punto de referencia compartido: un dataset público, anotado por profesionales del derecho mexicano, con metodología definida y validada por un comité independiente, sobre el cual cualquier sistema —comercial, académico o de código abierto— pueda evaluarse de forma transparente y reproducible.

§ 2 · PROPUESTA

¿Qué es BELMA?

BELMA se compone de cuatro elementos que operan en conjunto.

Primero, un dataset abierto: un conjunto de tareas legales mexicanas anotadas por profesionales del derecho, de acceso público para investigación y evaluación.

Segundo, una metodología reproducible: rúbricas y criterios de evaluación documentados de forma tal que cualquier equipo pueda correr el benchmark y verificar resultados de manera independiente.

Tercero, diversidad legal: múltiples áreas del derecho mexicano y un mix de tareas definidas por el comité, lo cual evita que el benchmark privilegie las fortalezas de un sistema específico.

Cuarto, una gobernanza independiente: las decisiones metodológicas las toma el Comité Técnico, integrado por miembros de academia, barra y práctica, en el cual la entidad que impulsa la iniciativa no ostenta mayoría ni poder de veto.

§ 3 · FASES DEL PROYECTO

Trayectoria de trabajo.

La construcción del benchmark se organiza en cuatro fases consecutivas.

I **Convocatoria y comité**
Inscripción abierta a participantes. Conformación del Comité Técnico Asesor con representación plural de instituciones académicas, barras de abogados y la práctica jurídica.

II **Definición metodológica**
El comité define la taxonomía de tareas, el esquema de anotación, las rúbricas de evaluación y la política de acceso al dataset. La metodología se publica para comentarios antes de cerrarse formalmente.

III **Construcción y anotación**
Sourcing de documentos a partir de fuentes públicas. Anotación distribuida con doble revisión y adjudicación. Reporte del *inter-annotator agreement* como métrica de calidad de las etiquetas.

IV **Validación y publicación**

Evaluación de modelos de referencia con resultados públicos. Lanzamiento del dataset, del paper técnico que documenta la metodología y de un leaderboard abierto a la industria y la academia.

§ 4 · PRINCIPIOS

Principios que rigen la iniciativa.

Neutralidad.
Las decisiones metodológicas, la selección de tareas y la curaduría del dataset son responsabilidad del Comité Técnico. La entidad que impulsa la iniciativa no ostenta mayoría ni poder de veto en dichas decisiones, y se compromete a respetar los acuerdos del comité aún cuando le resulten desfavorables.

Transparencia.
El dataset, la metodología, el código de evaluación y los resultados son públicos. La entidad que impulsa la iniciativa publicará sus propios resultados sin importar la posición que ocupe en el leaderboard, y se compromete a hacerlo antes de invitar a competidores comerciales a participar.

Rigor metodológico.
El proceso de anotación contempla doble revisión, adjudicación de discrepancias y reporte de *inter-annotator agreement*. La metodología es revisada por miembros externos del comité antes de la publicación del dataset en su versión 1.0.

§ 5 · PARTICIPACIÓN

¿A quién buscamos?

La convocatoria está abierta a cuatro perfiles complementarios. **Academia e investigación**: profesores, investigadores y estudiantes de posgrado en derecho o en ciencias de la computación con interés en metodologías de evaluación. **Despachos y litigantes**: abogados con práctica activa en cualquier área del derecho mexicano, quienes aportan criterio profesional sobre la dificultad y relevancia de cada tarea. **Equipos in-house**: áreas legales

corporativas que evalúan o utilizan herramientas de IA jurídica y que cuentan con perspectiva sobre los criterios reales de utilidad. **Estudiantes de derecho** de los últimos semestres con interés en investigación, tecnología legal o metodologías de evaluación.

¿Qué obtienes al participar?

Los anotadores principales y miembros del comité figuran como autores del paper técnico que acompañará el lanzamiento del benchmark. La participación queda registrada de forma pública en el sitio del proyecto y en las publicaciones derivadas, con la institución y rol del participante. Cada colaborador recibe un certificado de participación verificable. La inscripción al programa no compromete dedicación específica: el comité definirá los roles y el alcance esperado de cada participante en función de la fase del proyecto.

CÓMO SUMARSE

> Registrándose en el sitio del proyecto: belma.org.mx. La inscripción es gratuita y abierta a personas físicas y morales. Para preguntas sobre la iniciativa, sobre la integración del Comité Técnico o sobre eventuales colaboraciones institucionales, el contacto directo es el correo del proyecto.

DECLARACIÓN DE CONFLICTO DE INTERÉS

BELMA es impulsada por Temis AI, Inc., quien aporta infraestructura y secretaría técnica durante la fase de arranque del proyecto. Las decisiones metodológicas, la selección de tareas y la curaduría del dataset son responsabilidad exclusiva del Comité Técnico, en el cual Temis no ostenta mayoría ni poder de veto. El dataset, la metodología, el código de evaluación y los resultados —incluidos los de Temis— son públicos para toda la industria y la academia. La presente declaración se actualiza con cada versión del documento.