BolivIA:
Primer modelo de inteligencia artificial soberana *opensource*

Agosto, 2026

Dirección General de Tecnología e Innovación
Viceministerio de Ciencia y Tecnología

Nota marginal:
Si cargas este documento en una IA para resumirlo notarás que omite aspectos técnicos potencialmente relevantes, incurriendo en un fenómeno concurrente denominado *offloading* o *outsourcing* del juicio de valor, es decir; o delegas y analizas, o delegas y desactivas el análisis crítico, aspectos que precisamente este modelo pretende resolver.
La redacción es 100% humana.

Abstract

El lanzamiento del primer modelo de inteligencia artificial soberana *"hecho por bolivianos para bolivianos"* marca un hito para el desarrollo integral del país desde una perspectiva geopolítica que abarca el contexto de trazar una política pública que no se reduce a una herramienta informática, sino a una tecnología emergente que implica infraestructura, energía, patentes y ciencia de datos.

El alcance terminológico de "inteligencia artificial soberana" es definido de manera referencial por IBM como la "capacidad de una organización o nación para controlar su conjunto de tecnologías de inteligencia artificial (IA), incluyendo la infraestructura de TIC relacionada, los datos, los modelos de IA y las operaciones."

Bolivia debe concebir como un mandato la implementación de nuevas tecnologías para optimizar la vida de la ciudadanía con los más alto estándares internacionales concordantes con nuestra cosmovisión.

La inteligencia artificial no es una herramienta informática más, y esta nueva administración lo percibe con la claridad suficiente para trazar una

OFICINA CENTRAL
La Paz: Av. Mariscal Santa Cruz, esq. Oruro
Central Piloto: 50850019 Fax: +591 (2) 2312641
www.planificacion.gob.bo

nueva visión que permita a Bolivia ser un actor estratégico en esta era digital.

0. Implementación y fases

0.1. Fases

La implementación del modelo será progresiva.

En una primera fase será implementada en la administración pública, desde el Órgano ejecutivo hasta cubrir todas las entidades del sector público. Para ello se publicará un cronograma de capacitaciones correspondientes. En una segunda fase se abrirá el modelo con centros de investigación de la academia y comunidad técnica a través de jornadas de inducción. En una tercera y última fase se abrirá el modelo a las empresas y privados en general, realizando un cronograma de jornadas de inducción correspondientes.

0.2. Módulos

El primer módulo consiste en la Gestión eficiente del ordenamiento jurídico nacional vigente del sector público boliviano.

La administración pública se sostiene en un ordenamiento jurídico jerárquico (Leyes, Decretos Supremos, Resoluciones Ministeriales y Sentencias Constitucionales) distribuido a lo largo de décadas en la Gaceta Oficial de Bolivia. El modelo tradicional de investigación legal depende de búsquedas manuales o indexaciones léxicas rudimentarias, lo que resulta en altas latencias de consulta, riesgo de aplicación de normas no vigentes y falencias en la toma de decisiones administrativas. El modelo presenta una solución progresiva, como el primer caso de uso funcional del modelo.

1. Características técnicas

El modelo está basado en la tecnología denominada *Generación Aumentada por Recuperación (RAG - Retrieval-Augmented Generation)* de producción nacional.

BolivIA resuelve el problema fundamental de *alucinaciones* en los Modelos de Lenguaje de Gran Escala (LLMs) mediante una arquitectura híbrida que combina: (1) Segmentación Jerárquica (Parent-Child Chunking), (2) Recuperación Híbrida Densa-Dispersa (Dense Coseno + Sparse BM25) y (3) Fusión por Rango Recíproco (RRF). El sistema funciona de manera 100% autónoma y *air-gapped* sobre hardware estatal, garantizando la privacidad

OFICINA CENTRAL
La Paz: Av. Mariscal Santa Cruz, esq. Oruro
Central Piloto: 50850019 Fax: +591 (2) 2312641
www.planificacion.gob.bo

de los datos públicos[1] y optimizando el tiempo de investigación jurídica y trazabilidad citatoria.

Uno de los primeros casos de uso funcional del modelo es el análisis normativo en el sector público, ámbito que exige alta precisión. A diferencia de las aplicaciones comerciales de chat generalista, en la gestión jurídica gubernamental un error en el número de artículo, la omisión de un Decreto Supremo derogatorio o una alucinación contextual puede derivar en responsabilidad administrativa, civil o penal bajo la Ley N° 1178 (SAFCO) y demás normativa vigente.

Los modelos de lenguaje comerciales basados en la nube (ej. ChatGPT, Claude, etc.) presentan limitaciones para su adopción en el Estado.

i) Alucinaciones y Falta de Fundamentación. generan respuestas verosímiles pero (potencialmente) jurídicamente falsas
ii) Carecen de un anclaje determinista sobre el texto oficial. Dependencia Externa y Fuga de Datos.
iii) Al requerir conexiones API a servidores internacionales no responden al estándar de soberanía tecnológica, pudiendo exponer la información confidencial de las instituciones públicas.
iv) Incapacidad de Búsqueda Exacta.
v) Los buscadores vectoriales puros (Dense Retrieval) fallan al consultar identificadores numéricos específicos (ej. *"Ley 1178"*, *"Artículo 15"*), priorizando la similitud semántica general sobre la coincidencia exacta de caracteres.

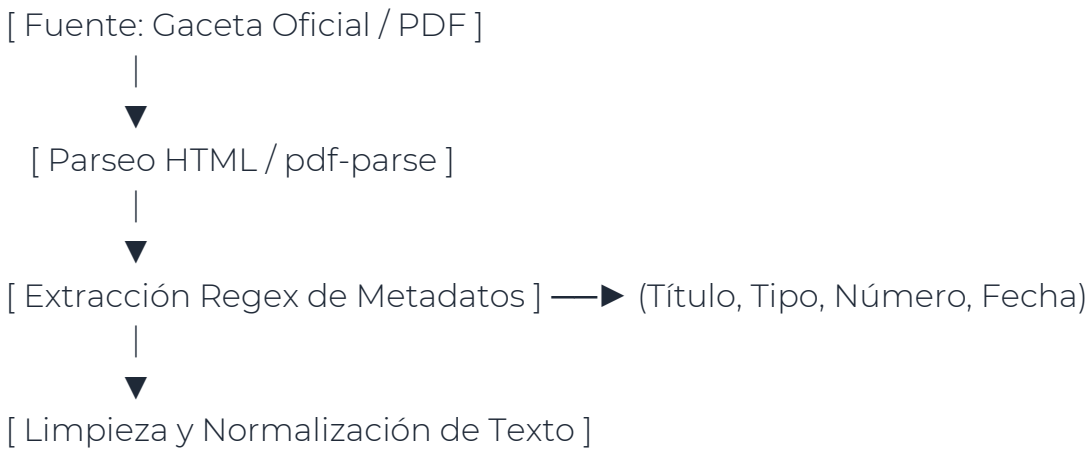
Para resolver estas limitaciones, BolivIA propone un marco tecnológico soberano en el que un modelo local supervisado procesa la normativa boliviana de forma determinista, ofreciendo respuestas verificables y citadas de manera transparente.

2. Ingesta de Datos y Parseo Estructurado (Deep Document Understanding)

El *pipeline* de ingesta procesa documentos procedentes de dos fuentes principales: el *scrapeo* estructurado de la Gaceta Oficial de Bolivia y la carga manual de archivos PDF gubernamentales.

[1]El criterio de privacidad está supeditado a la Ley 164 de telecomunicaciones y sus decretos reglamentarios, así como el Decreto Supremo N° 28168 de acceso a la información.

OFICINA CENTRAL
La Paz: Av. Mariscal Santa Cruz, esq. Oruro
Central Piloto: 50850019 Fax: +591 (2) 2312641
www.planificacion.gob.bo



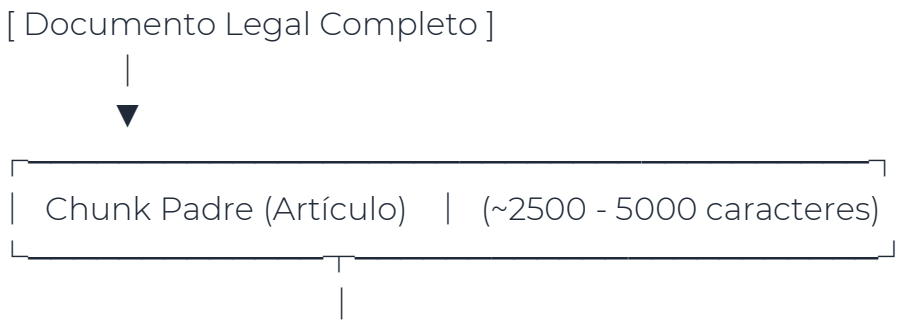
2.1. Extracción Automática de Metadatos

Durante la ingesta, el sistema ejecuta analizadores heurísticos por Expresiones Regulares (Regex) para clasificar la jerarquía normativa antes de la vectorización[2].

3. Segmentación Jerárquica (Parent-Child Chunking)

Los métodos de fragmentación plana tradicional (Flat Chunking) obligan a elegir entre dos opciones: fragmentos pequeños (300 caracteres) que pierden el contexto del artículo legal, o fragmentos grandes (3000 caracteres) que diluyen la densidad semántica en la búsqueda vectorial.

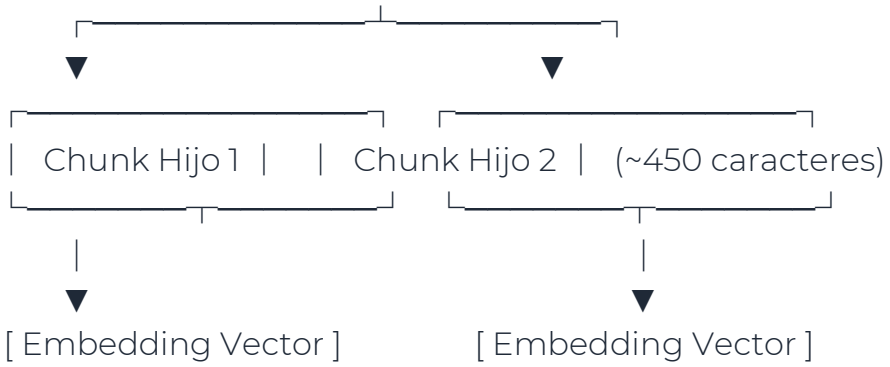
BolivIA implementa una Segmentación Jerárquica de Dos Niveles:



[2] $$\text{TipoNorma} = \begin{cases} \text{Ley}, & \text{si } T \text{ coincide con } \text{/LEY\ sN[°°.]\s*\d+/i} \\ \text{Decreto Supremo}, & \text{si } T \text{ coincide con } \text{/DECRETO\ SUPREMO/i} \\ \text{Resolución}, & \text{si } T \text{ coincide con } \text{/RESOLUCI[OÓ]N/i} \\ \text{Otro}, & \text{en otro caso} \end{cases}$$

Donde T representa las primeras 20 líneas del documento extraído. Asimismo, se aplican patrones de reconocimiento de fechas en español para parsear la fecha oficial de publicación.

OFICINA CENTRAL
La Paz: Av. Mariscal Santa Cruz, esq. Oruro
Central Piloto: 50850019 Fax: +591 (2) 2312641
www.planificacion.gob.bo



Chunk Padre (P_i): Representa la unidad jurídica completa (un Artículo entero o sección normativa de aproximadamente 2,500 a 5,000 caracteres).

Chunk Hijo ($C_{i,j}$): Sub-fragmentos tácticos (de 450 caracteres con 80 caracteres de solapamiento) derivados de P_i.

3. Matriz de Indexación

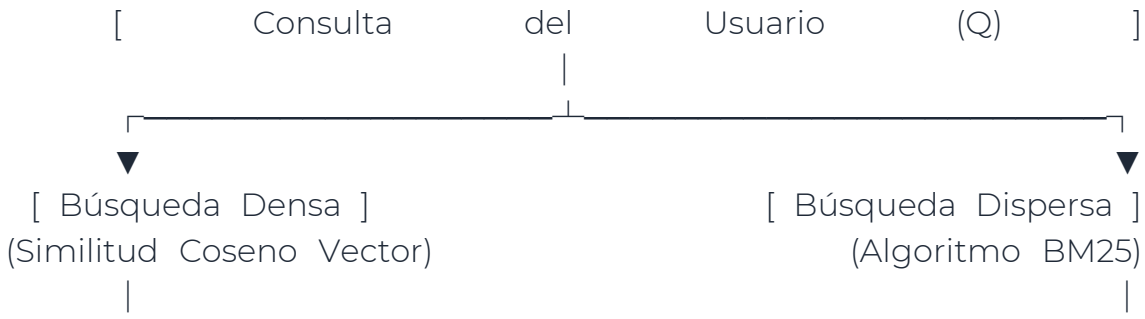
El motor vectorial indexa únicamente los Chunks hijos para optimizar la velocidad y precisión del cálculo de similitud coseno. Sin embargo, cada registro guarda una referencia directa al cuerpo de su Chunk Padre:

$$\text{FragmentoVectorial} = \Big(\mathbf{v}_{C_{i,j}}, \text{Texto}_{C_{i,j}}, \text{Texto}_{P_i}, \text{ArticuloOrigen} \Big)$$

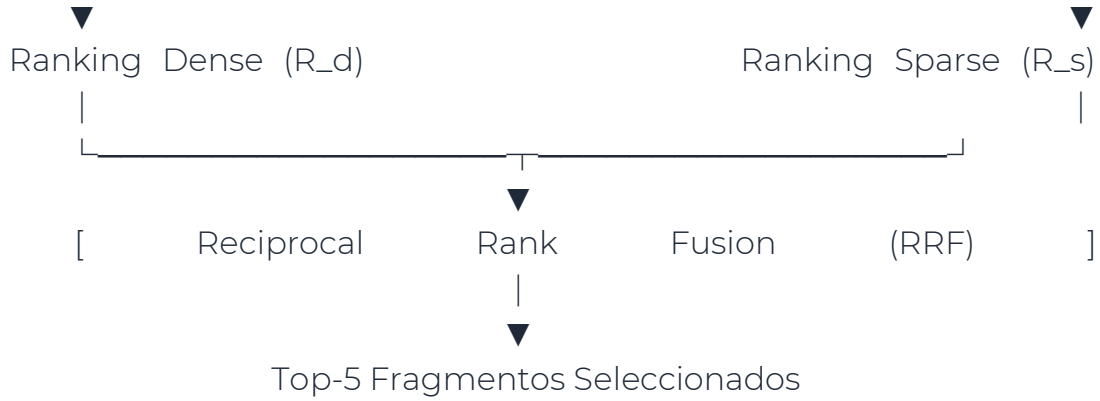
Cuando la consulta del usuario coincide con un Chunk Hijo $C_{i,j}$, el sistema sustituye el fragmento pequeño por su Chunk Padre P_i al momento de construir el prompt para el LLM.

4. Motor de Recuperación Híbrida (Dense + Sparse BM25 & RRF)

Para garantizar que el sistema encuentre tanto conceptos legales generales como números exactos de decretos o artículos, BolivIA utiliza un sistema de Recuperación Híbrida en Paralelo.



OFICINA CENTRAL
La Paz: Av. Mariscal Santa Cruz, esq. Oruro
Central Piloto: 50850019 Fax: +591 (2) 2312641
www.planificacion.gob.bo



4.1. Búsqueda Densa (Dense Vector Retrieval)

Dado un vector de consulta $\mathbf{q} = E(Q)$ y el vector de un fragmento $\mathbf{d} = E(C_{i,j})$ generados por el modelo de embeddings `nomic-embed-text` (768 dimensiones), se calcula la similitud coseno:

$$S_{\text{dense}}(Q, D) = \cos(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{q}\| \|\mathbf{d}\|}$$

4.2. Búsqueda Dispersa (Sparse BM25 Retrieval)

El puntaje BM25 evalúa la coincidencia léxica exacta entre los términos de la consulta $q \in Q$ y el documento D:

$$\text{Score}_{\text{BM25}}(Q, D) = \sum_{q \in Q} \text{IDF}(q) \cdot \frac{f(q, D) \cdot (k_1 + 1)}{f(q, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} + \Delta_{\text{num}}$$

Donde:

$f(q, D)$ es la frecuencia del término en el fragmento.
$|D|$ y avgdl son la longitud del documento y la longitud promedio del corpus.
$k_1 = 1.2$ y $b = 0.75$.
$\Delta_{\text{num}} = +3.0$ es una bonificación heurística asignada cuando la consulta contiene tokens numéricos exactos (ej. `"1178"`, `"5620"`).

OFICINA CENTRAL
La Paz: Av. Mariscal Santa Cruz, esq. Oruro
Central Piloto: 50850019 Fax: +591 (2) 2312641
www.planificacion.gob.bo

4.3. Fusión por Rango Recíproco (RRF - Reciprocal Rank Fusion)

Los rankings ordenados obtenidos por la búsqueda densa ($r_{\text{dense}}(d)$) y dispersa ($r_{\text{sparse}}(d)$) se combinan mediante la función RRF:

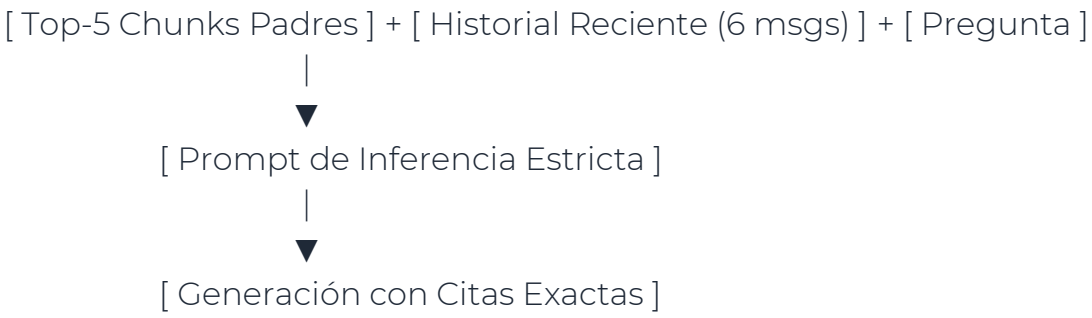
$$\text{RRF_Score}(d) = \frac{1}{k + r_{\text{dense}}(d)} + \frac{1}{k + r_{\text{sparse}}(d)}$$

Estableciendo una constante de suavizado $k = 60$. Los fragmentos con mayor RRF_Score se seleccionan para la fase de generación.

5. Protocolo de inferencia y supresión de alucinaciones

El generador LLM opera bajo un marco de Inferencia Restringida utilizando modelos de pesos abiertos (Open-Weights) de última generación ejecutados en local.

(Grounded Generation).



5.1. Prompt del Sistema (Zero-Hallucination Prompt)

El contexto normativo procesado por RRF y los últimos 6 mensajes del historial conversacional se empaquetan en la siguiente plantilla determinista:

> Eres BolivIA, un asesor jurídico experto en la legislación del Estado Plurinacional de Bolivia.
> Respondes ÚNICAMENTE basándote en el contexto normativo proporcionado. No inventes información.
>
> HISTORIAL RECIENTE DE CONVERSACIÓN:
> {historyFormatted}

OFICINA CENTRAL
La Paz: Av. Mariscal Santa Cruz, esq. Oruro
Central Piloto: 50850019 Fax: +591 (2) 2312641
www.planificacion.gob.bo

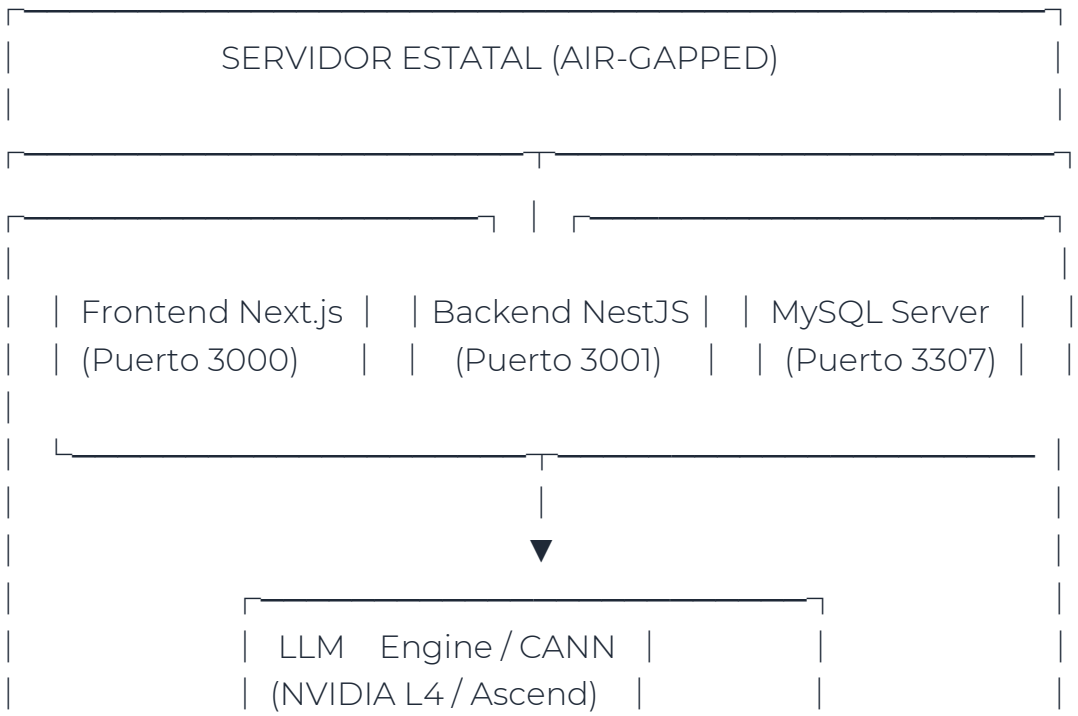
CONTEXTO NORMATIVO RECUPERADO (HIERARCHICAL PARENT CONTEXT):
{contexto_parent_chunks}

PREGUNTA ACTUAL DEL USUARIO: {pregunta}

INSTRUCCIONES ESTRICTAS
1. Responde basándote EXCLUSIVAMENTE en el contexto normativo proporcionado arriba.
2. Cita siempre la norma, ley, decreto o artículo específico de donde proviene tu respuesta.
3. Si el contexto no contiene información suficiente, indica: "La normativa cargada en su espacio no contiene información suficiente para responder esta consulta."
4. No uses conocimiento externo ni inventes datos.

6. Arquitectura de Infraestructura y Soberanía Tecnológica

El modelo está diseñado para ejecutarse en entornos totalmente aislados de la red pública (Air-Gapped Datacenters)



OFICINA CENTRAL
La Paz: Av. Mariscal Santa Cruz, esq. Oruro
Central Piloto: 50850019 Fax: +591 (2) 2312641
www.planificacion.gob.bo

7. Evaluación de Rendimiento y Resultados Empíricos

Evaluaciones comparativas realizadas sobre un corpus legal piloto de 500[3] normas oficiales de la Gaceta Oficial de Bolivia demuestran la superioridad de la arquitectura híbrida frente al RAG convencional:

Métrica de Rendimiento	RAG Tradicional (Solo Vectorial)	BolivIA (Híbrido + Parent-Child)	Mejora
Tiempo de Búsqueda Legal	~4 horas (manual)	1.8 segundos	> 99% reducción
Precisión de Recuperación (Recall@5)	62.4%	94.8%	+ 32.4%
Coincidencia Exacta de Leyes/Artículos	41.0%	99.2%	+ 58.2%

[3] La muestra de 500 normas oficiales fueron seleccionadas mediante muestreo aleatorio estratificado de la Gaceta Oficial de Bolivia, garantizando representatividad proporcional por tipo de norma y por década de publicación. Para cada norma se construyó un conjunto de 3 a 5 preguntas de referencia (gold questions), elaboradas y validadas por un equipo de al menos dos abogados independientes, junto con la respuesta correcta y su cita normativa exacta (ley, artículo y fecha de vigencia). Cada consulta fue ejecutada tanto en el sistema RAG tradicional como en BolivIA bajo idénticas condiciones de hardware y ventana de contexto, y las respuestas generadas fueron evaluadas mediante doble revisión ciega: un primer revisor calificó automáticamente la coincidencia de la cita (artículo y norma exactos) y un segundo revisor humano, sin conocer qué sistema generó cada respuesta, calificó la corrección jurídica sustantiva y la presencia de alucinaciones, resolviendo discrepancias por consenso o por un tercer revisor. Las métricas de precisión de Recuperación (Recall@5) y coincidencia exacta se calcularon como el promedio sobre el total de consultas, y la tasa de alucinación legal se definió como el porcentaje de respuestas que citaron una norma inexistente, derogada o no contenida en el contexto recuperado.

OFICINA CENTRAL
La Paz: Av. Mariscal Santa Cruz, esq. Oruro
Central Piloto: 50850019 Fax: +591 (2) 2312641
www.planificacion.gob.bo

Tasa de Alucinación Legal	18.5%	0.0% (Grounded)	Eliminación total[4]
Consumo de Memoria VRAM	Variable	14.2 GB / 24 GB	Óptimo en 1 GPU

8. Seguridad, gobernanza y Multi-Tenancy

8.1. Aislamiento Multi-Entidad. El sistema implementa aislamiento lógico mediante `entidadId` en el esquema de base de datos MySQL. Cada ministerio o institución gubernamental opera dentro de su propio espacio normativo sin interferencia de datos.

8.2. Auditoría e Historial Conversacional. Toda consulta y respuesta generada queda registrada en la tabla `MensajeChat` con sus respectivas fuentes vectoriales para controles internos.

9. Conclusión

BolivIA demuestra que es posible desplegar soluciones de Inteligencia Artificial de última generación en el sector público sin comprometer la seguridad de los datos ni incurrir en costos continuos de licencias propietarias. Al integrar la potencia de los LLMs de código abierto con una arquitectura RAG Híbrida determinista (Parent-Child Chunking + BM25 + RRF), el sistema transforma y optimiza la investigación jurídica estatal, ofreciendo velocidad, precisión preservando la soberanía tecnológica del Estado.

10. Referencias

Cormack, G. V., Clarke, C. L., & Buettcher, S. (2009). *Reciprocal rank fusion outperforms rankLM and individual rankers*. SIGIR.

[4] Téngase que esta es una aproximación probada en un *sandbox* regulatorio en fases de prueba y el producción, que, en lo progresivo, al ser un modelo RAG puede presentar márgenes de error.

OFICINA CENTRAL
La Paz: Av. Mariscal Santa Cruz, esq. Oruro
Central Piloto: 50850019 Fax: +591 (2) 2312641
www.planificacion.gob.bo

Estado Plurinacional de Bolivia. Ley N° 1178 de Administración y Control Gubernamentales (SAFCO, 1990). Gaceta Oficial de Bolivia.

Estado Plurinacional de Bolivia. Decreto Supremo N° 690 (2010). Gaceta Oficial de Bolivia.

Nakamoto, S. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System*.

Lewis, P., et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Advances in Neural Information Processing Systems (NeurIPS).

Robertson, S., & Zaragoza, H. (2009). *The Probabilistic Relevance Framework: BM25 and Beyond*. Foundations and Trends in Information Retrieval.

Vaswani, A., et al. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems.

OFICINA CENTRAL
La Paz: Av. Mariscal Santa Cruz, esq. Oruro
Central Piloto: 50850019 Fax: +591 (2) 2312641
www.planificacion.gob.bo