

Un RAG portable sobre SUDOCU: consulta conversacional de normativa universitaria mediante estados de diálogo con procedencia

Juan Manuel Fernández¹, Francisco Tamasi¹, Sebastián Marchetti¹, Mario Oloriz¹, and Marcelo Errecalde²

¹ LICDIA, Universidad Nacional de Luján, Luján, Argentina

² LIDIC, Universidad Nacional de San Luis, San Luis, Argentina
{jmfernandez,ftamasi,smarchetti,moloriz}@unlu.edu.ar, merreca@unsl.edu.ar

Resumen La normativa de las universidades nacionales argentinas es pública, pero los portales que la publican solo admiten búsqueda por atributos del acto: localizar qué norma regula una situación exige conocer de antemano el acto que la contiene. Se presenta un sistema de generación aumentada por recuperación (RAG) sobre el módulo de publicación documental de SUDOCU, desplegado sobre los 21.344 actos administrativos de la Universidad Nacional de Luján, segmentados por unidad normativa en 151.696 fragmentos. La recuperación combina *embeddings* multilingües con un índice léxico que preserva los identificadores normativos, fusionados por posición. Sobre esa base, el sistema mantiene un estado de diálogo con dos *slots* —la entidad en discusión y los actos mencionados— cuyo peso en la recuperación lo determina la *procedencia* de cada valor: lo que el sistema infiere pondera débilmente; lo que el usuario fija, visible y editable en la interfaz, prevalece y no se sobrescribe. En la evaluación, el anclaje resuelve por completo las consultas por identificador (R@5 de 1,00), el procesamiento conversacional lleva la repregunta elíptica de 0,12 a 0,93 y el 97,0% de los actos citados tiene correspondencia verificada con lo recuperado. La portabilidad se verificó construyendo una segunda instancia sobre el portal de la Universidad Nacional de San Luis, con el mismo código y solo configuración.

Keywords: RAG, SUDOCU, Estado de diálogo, Normativa universitaria.

1. Introducción

En Argentina, las universidades nacionales producen su normativa mediante actos administrativos —resoluciones, disposiciones, ordenanzas— cuya publicación en línea avanzó de manera desigual: no existe un canal único y cada institución combina digestos propios, boletines oficiales y sistemas de gestión documental. En ese escenario, el Sistema Único Documental (SUDOCU), desarrollado por la Universidad Nacional de General Sarmiento en el marco del consorcio SIU, se consolidó como plataforma de gestión y publicación documental del sistema universitario: 37 instituciones utilizan su solución de expediente

electrónico integrado³, y su módulo de publicación expone los actos administrativos en un portal público. La normativa, sin embargo, sigue exigiendo un conocimiento previo que la mayoría de sus destinatarios no tiene: para encontrar una resolución hay que saber qué resolución se busca. Los portales de publicación documental ofrecen búsqueda por número, tipo de acto y órgano emisor: la búsqueda opera sobre atributos del acto y no sobre su contenido, de modo que localizar qué norma regula una situación concreta exige conocer de antemano el acto que la contiene.

Este trabajo presenta un sistema de generación aumentada por recuperación (RAG) [14]⁴ sobre el módulo de publicación documental de SUDOCU, implementado y evaluado sobre la totalidad de los actos publicados por la Universidad Nacional de Luján (21.344 documentos) y desplegado además sobre más de 26.000 documentos de una segunda universidad nacional (San Luis). El sistema responde en lenguaje natural y cita los fragmentos normativos que sustentan sus afirmaciones, con trazabilidad hasta el documento oficial publicado.

A su vez, la consulta de normativa es conversacional —se pregunta por un tema, se identifica un acto y se repregunta sobre su articulado, con repreguntas elípticas que omiten el sujeto— y los sistemas de diálogo orientados a tareas abordan esa dependencia manteniendo un estado que resume la conversación [23,24]. Este trabajo adopta ese mecanismo con una función distinta: el estado no completa los campos de una consulta sobre un esquema cerrado, sino que *pondera* la recuperación sobre un corpus documental abierto, apoyada en los metadatos estructurados que el sistema de origen provee.

El aporte central es el criterio que gobierna esa ponderación. En la literatura, la aplicación dura o blanda de una restricción depende de la incertidumbre del modelo, del tipo de atributo o de la arquitectura del sistema. Aquí depende de la *procedencia* del valor: una entidad inferida por el sistema refuerza débilmente la recuperación; una fijada por el usuario —que ve el estado y puede editarlo— recibe mayor peso y no se sobrescribe. La corrección del usuario se trata como una observación de mayor confianza que la inferencia del propio sistema.

Las contribuciones son: (i) un sistema desplegado de consulta conversacional de normativa universitaria, con trazabilidad desde las afirmaciones hasta el documento publicado, fragmentación por unidad normativa según la forma jurídica del acto administrativo argentino, y operación —catálogo, descarga, vectorización, indexación— íntegramente autoadministrada; (ii) un estado de diálogo que conserva los referentes de la conversación, expone la interpretación del sistema y admite corrección directa, donde la *procedencia* de cada valor determina su peso en la recuperación; y (iii) portabilidad verificada: una segunda instancia opera sobre el portal SUDOCU de otra universidad nacional con el mismo código y solo configuración (Sección 4).

El resto del artículo presenta los antecedentes (Sección 2), la metodología —corpus, sistema y protocolo de evaluación— (Sección 3), los experimentos y sus resultados (Sección 4) y las conclusiones (Sección 5).

³ <https://sudocu.dev/>, consultado en agosto de 2026.

⁴ Instancia de demostración: <https://licdia.unlu.edu.ar/rag-unlu>.

2. Antecedentes

La aplicación de IA en la educación superior crece aceleradamente, pero con una asimetría documentada: solo el 11 % de los estudios se orienta a la gestión y administración, frente al 72 % centrado en estudiantes [5], y buena parte de la literatura describe potencialidades antes que implementaciones consolidadas [8]. Los antecedentes funcionales más cercanos son asistentes de consulta de rutina sobre políticas académicas [18], sin recuperación sobre el corpus normativo ni trazabilidad a la norma publicada. Este trabajo se ubica en esa vacancia —particularmente aguda en el contexto latinoamericano— con un sistema desplegado de consulta normativa para la gestión universitaria.

La generación aumentada por recuperación acopla un modelo de lenguaje con un índice de búsqueda: la respuesta se redacta condicionada a fragmentos recuperados del corpus, lo que reduce las alucinaciones [20] y habilita la cita de fuentes [7,14]. El enfoque se aplicó a dominios legales y regulatorios, donde la exigencia de respaldo documental es mayor que en un dominio abierto: la evaluación empírica de herramientas comerciales de investigación jurídica asistida registra tasas de alucinación significativas incluso con recuperación aumentada [16].

En la dimensión conversacional, el seguimiento del estado de diálogo es estándar en sistemas orientados a tareas [23,24], donde los *slots* (ranuras) completan una consulta a una base con esquema cerrado y conocido. Varios trabajos recientes lo usan, en cambio, para restringir la recuperación. ShopTalk [17] convierte el estado de diálogo en una combinación de consulta textual y restricciones de faceta, y ya combina un régimen duro y uno blando según el atributo esté o no anclado al catálogo. Thulke et al. [21] usan el contexto del diálogo para acotar jerárquicamente el espacio de candidatos antes del *ranking*. RA-Rec [11] nombra explícitamente restricciones duras y blandas en el esquema del estado, aunque las resuelve colapsadas en una consulta en lenguaje natural.

El sistema que se presenta en este trabajo comparte con estos antecedentes el uso del estado como señal de recuperación. Se diferencia en dos puntos: el corpus es documental y abierto —no un catálogo de ítems ni un esquema tipo Multi-WOZ [2]— y, sobre todo, el régimen de aplicación no lo decide la incertidumbre del modelo ni el tipo de atributo, sino la procedencia del valor.

Por último, que el usuario pueda inspeccionar y corregir lo que un sistema infirió es un principio establecido, de los sistemas *penetrables* de Koenemann y Belkin [12] al ciclo de explicación y corrección de Kulesza et al. [13]. En diálogo orientado a tareas, SOVITE [15] muestra los *slots* detectados y permite repararlos por manipulación directa; en recuperación, propuestas recientes plantean exponer y dejar editar la intención inferida [1].

Los antecedentes relevados aportan estas piezas por separado: el estado de diálogo como señal de recuperación, por un lado, y la transparencia y corrección del estado inferido como principio de diseño, por el otro. No se registra, en cambio, un sistema que las combine —exponer el estado, admitir su edición y hacer de la *procedencia* del valor el criterio que gobierna el régimen de recuperación— ni una aplicación de ese esquema a un corpus documental abierto de normativa institucional en un sistema en operación en Argentina.

3. Metodología

3.1. Construcción del corpus

La Figura 1 resume el flujo desde el portal hasta los artefactos de búsqueda. El módulo de publicación documental (MPD) de SUDOCU expone los actos publicados; la etapa de ingesta consume la API REST del portal —la misma que utiliza su interfaz— y produce, por un lado, el *catálogo*: los metadatos desglosados de cada acto (tipo, código, número, fecha, título, órgano emisor), su identificador interno y una URL permanente al PDF oficial; por el otro, los PDF. La completitud se verifica contra el total que el portal declara por carpeta: 21.344 actos en once carpetas, con cobertura desde abril de 2024.

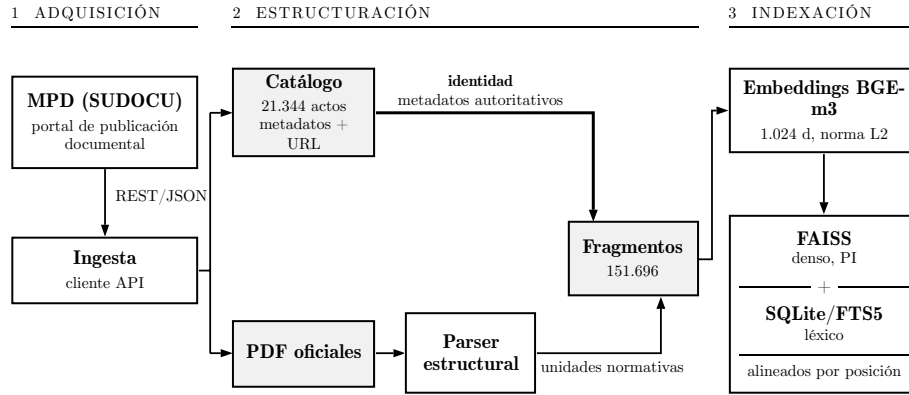


Figura 1. Arquitectura de construcción del corpus.

El parser convierte cada PDF a texto estructurado y lo segmenta por unidad normativa —Visto, Considerando, parte resolutive, artículos, anexos—: la unidad de recuperación coincide con la unidad citable, de modo que una respuesta puede referir el artículo 2 de una disposición y enlazar exactamente ese texto. Cada fragmento porta la identidad y la trazabilidad del acto —código, número, año, fecha, identificadores del portal y URL permanente—, tomadas del catálogo y no del PDF.

3.2. Vectorización e indexación

Los *embeddings* se calculan con BGE-m3 [3], modelo multilingüe de recuperación entrenado contrastivamente sobre pares masivos, con soporte de representación densa, dispersa y multivector; se emplea la representación densa, de 1.024 dimensiones y normalizada L2, de modo que el producto interno equivale a la similitud coseno y el índice puede ser exacto y plano. Se vectorizan los 151.696 fragmentos del corpus (7,1 por documento; 55.929 artículos). La ventana

de 8.192 tokens admite fragmentos largos sin truncamiento. Cada fragmento se codifica con un encabezado sintético —título del acto y cita citable, antepuestos al cuerpo—: un condicionamiento de contexto documental sobre la representación local que mitiga la pérdida de contexto inherente a la fragmentación, en línea con la concatenación de título y pasaje de los recuperadores densos [10]. Los artefactos de búsqueda son un índice FAISS [6] de producto interno y una base SQLite con el texto, los metadatos y un índice léxico FTS5⁵, alineados por posición y verificados tras cada reconstrucción; el reemplazo es atómico y la API recarga el índice sin interrumpir el servicio. La operación completa —catálogo, descarga, vectorización, indexación— se ejecuta desde el panel de administración del propio sistema, con registro de corridas, y la actualización periódica reutiliza el mismo camino de forma incremental.

3.3. Recuperación híbrida

La recuperación combina la búsqueda por *embeddings* de la Sección 3.2 con recuperación léxica: una implementación propia de BM25 [19] ($k_1=1,5$, $b=0,75$) cuyo tokenizador *preserva los identificadores normativos* —expresiones como RESHCS-LUJ: 0000042-24, 528/2025 o DISPCD-CB se mantienen como unidades— porque son exactamente lo que los *embeddings* no distinguen. Los dos *rankings* se fusionan por *reciprocal rank fusion* (RRF) [4], que asigna a cada documento la suma de los recíprocos de su posición en cada *ranking*:

$$\text{RRF}(d) = \sum_{r \in \{\text{emb}, \text{BM25}\}} \frac{1}{K + \text{rango}_r(d)}, \quad K = 60, \quad (1)$$

donde $\text{rango}_r(d)$ es la posición del documento d en el *ranking* r y la constante K amortigua la diferencia entre las primeras posiciones. Fusionar por posiciones, y no por puntajes, evita calibrar magnitudes de naturaleza distinta —similitud coseno y puntaje BM25—. Cuando la consulta nombra un acto concreto, la búsqueda se restringe a los documentos que llevan ese número. La fusión por sí sola no lo garantiza: un documento con posiciones intermedias en ambos *rankings* supera a uno que encabeza uno solo. Además, el número solo no alcanza, porque cada tipo de acto lleva numeración propia y un mismo número existe en decenas de códigos distintos.

3.4. Estado de diálogo con procedencia

En cada turno, una única pasada de un modelo de lenguaje sobre los últimos intercambios produce conjuntamente una consulta autocontenida y una salida estructurada con los referentes detectados. La consulta reescrita repone la información omitida en las repreguntas elípticas, mientras que esos mismos referentes actualizan los *slots* del estado de diálogo y permanecen disponibles para su inspección, corrección y ponderación durante los turnos siguientes.

⁵ <https://sqlite.org/fts5.html>, consultado en agosto de 2026.

El estado mantiene dos *slots*. El primero, *entidad*, registra el sujeto de la conversación —una persona, una carrera, una dependencia, un órgano—, con cardinalidad uno y semántica de reemplazo. El segundo, *actos*, acumula el conjunto de actos administrativos mencionados. El criterio de admisión fue restrictivo: un *slot* se justifica únicamente si transporta entre turnos información que el texto de la consulta actual no contiene.

Los *slots* se muestran sobre el campo de escritura desde la primera recuperación y permanecen visibles durante la conversación. El estado es editable: quien consulta puede corregir la entidad, descartar un acto o reincorporarlo. Lo descartado permanece visible, atenuado, dado que la inspección de lo inferido es condición para su control. Cuando una entidad fijada no aparece en la consulta con la que se buscará, el sistema declara la discrepancia en la interfaz en lugar de resolverla de manera implícita.

Cada valor del estado registra, además, su *procedencia*, de la que se deriva su peso en la recuperación: un valor inferido por el modelo a partir de la conversación pesa 0,5; uno fijado explícitamente por quien consulta pesa 1,0 y deja de ser sobrescrito por la inferencia en los turnos siguientes; uno descartado pesa 0. El peso entra como bonificación aditiva sobre el puntaje fusionado:

$$\text{puntaje}(d) = \text{RRF}(d) + \sum_{s \in S} w_s \cdot \mathbb{K}[d \models s] \cdot u \quad (2)$$

donde S es el conjunto de *slots* activos, w_s el peso derivado de la procedencia y $u = 1/(K+1)$ la unidad de bonificación, equivalente a encabezar uno de los dos *rankings*. Expresarla en unidades de RRF evita una constante arbitraria y acota el efecto: un *slot* desplaza el ordenamiento, pero no sustituye a la relevancia, que aporta por ambos *rankings*. La corrección del usuario opera así como una observación de mayor confianza que la inferencia del propio sistema: duplica el peso del valor en la Ecuación 2 y fija su persistencia.

Ningún *slot* excluye documentos: un valor mal inferido degrada el ordenamiento pero no puede anular el conjunto de resultados, el modo de falla característico de los filtros estrictos.

El estado cumple así tres funciones: conserva los referentes que resuelven la elipsis, expone la interpretación que el sistema mantiene y admite su corrección directa; la procedencia de cada valor determina su peso y su persistencia en la recuperación.

3.5. Generación

La respuesta se redacta con un modelo de lenguaje que recibe la consulta, el contexto recuperado —fragmentos con su cita— y la instrucción de respaldar cada afirmación con el identificador del acto correspondiente, sin afirmar nada que el contexto no contenga. El componente es intercambiable: el sistema admite cualquier *endpoint* compatible con la API de OpenAI —servicios en la nube o modelos de pesos abiertos servidos localmente (vLLM, Ollama)— y se configura desde el panel de administración, sin reiniciar el servicio. Sobre la misma recuperación, el sistema expone modos de salida invocables por comandos

(/lista, /ficha, /comparar, /novedades, /resumen) que reconfiguran la selección de contexto y la instrucción de generación: inventarios por tema, dossier de un acto —incluida la relación inversa de qué actos lo mencionan, que el portal no ofrece—, comparación entre dos actos, novedades por ventana temporal y síntesis narrada.

3.6. Protocolo de evaluación

Ante la ausencia de un conjunto anotado sobre este corpus, la recuperación se evalúa con consultas sintéticas construidas por muestreo estratificado por sección del portal y tipo de acto, de tres clases: por *identificador* ($n = 30$; plantillas deterministas sobre el código del acto), por *contenido* ($n = 131$; generadas con un modelo de lenguaje a partir de un fragmento, sin nombrar el número del acto, con un control de calidad automático —un segundo pase que juzga si el fragmento responde la consulta— que descartó el 34,5 %) y *conversacionales* de dos turnos ($n = 58$; el segundo turno es una repregunta elíptica). La relevancia se decide a nivel de acto —el sistema cita y enlaza actos completos— y se reportan la cobertura ($\text{Recall}@k$) y la precisión del ordenamiento ($\text{MRR}@10$ [22], $\text{nDCG}@10$ [9]). La fidelidad de las citas se verifica de forma determinista: todo acto citado en una respuesta debe estar entre lo recuperado, y la atribución a una persona nombrada debe estar contenida en el fragmento citado⁶.

El juicio humano complementa las métricas automáticas: cinco evaluadores —dos estudiantes, un docente, un evaluador externo a la institución y un integrante del área de gestión— completan un cuestionario único con quince respuestas fijas del sistema —idénticas entre jueces, lo que permite reportar acuerdo—, puntuadas de 1 a 5 en utilidad, fidelidad, completitud y claridad⁷.

4. Experimentos y resultados

La evaluación automática se centra en la recuperación: la Tabla 1 compara el sistema completo con los enfoques denso, léxico e híbrido sin procesamiento conversacional.

Tabla 1. Recuperación por tipo de consulta. En las conversacionales, todas las configuraciones salvo el sistema completo operan sobre el segundo turno crudo.

Configuración	Identificador			Contenido			Conversacional		
	R@5	MRR	nDCG	R@5	MRR	nDCG	R@5	MRR	nDCG
Solo <i>embeddings</i>	0.10	0.08	0.08	0.78	0.64	0.69	0.10	0.08	0.09
Solo BM25	0.83	0.68	0.76	0.79	0.68	0.71	0.09	0.08	0.08
Híbrido	1.00	1.00	1.00	0.82	0.72	0.75	0.12	0.07	0.09
Híbrido + estado	1.00	1.00	1.00	0.82	0.72	0.75	0.93	0.88	0.90

⁶ Tanto la generación de consultas sintéticas como la de respuestas usó `gpt-4o-mini` a temperatura 0.

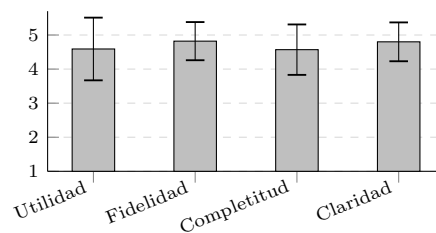
⁷ El cuestionario y los ítems evaluados están disponibles en el repositorio del sistema: <https://github.com/jumafernandez/rag-unlu> (directorio `evaluacion/`).

Ninguno de los dos componentes alcanza por sí solo: los *embeddings* no distinguen identificadores normativos y BM25, aun preservándolos, no ordena la semántica de las consultas por contenido tan bien como la fusión; el híbrido resuelve por completo las consultas por identificador y conserva la ventaja en las de contenido. El resultado perfecto de la primera columna se explica por el anclaje de identificadores que integran las configuraciones híbridas: cuando la consulta nombra un acto, la búsqueda se restringe a los documentos que llevan ese identificador, y el acto correcto encabeza el *ranking* con independencia de la fusión; sin esa restricción, la fusión por posiciones no lo garantiza. En las consultas conversacionales ancladas en un acto, esa misma restricción se habilita porque el estado repone el identificador omitido en la consulta reescrita: la composición de ambos mecanismos explica el 1,00 de esa población. La columna conversacional muestra el papel del estado: sobre el segundo turno crudo ninguna configuración recupera (R@5 de 0,09 a 0,12) y el sistema completo alcanza 0,93, con la reescritura condicionada al historial como mecanismo determinante; los *slots*, cuyos valores ya están repuestos en la consulta reescrita, conservan su función de corrección con prioridad garantizada por la procedencia. Un *cross-encoder* (BGE-reranker-v2-m3) reordenando el top-50 del híbrido no mejora las métricas (R@1 global de 0,68 contra 0,70) y empeora los identificadores: el ordenamiento no es el cuello de botella de este dominio. La verificación determinista de citas sobre 60 respuestas ancla el 97,0 % de los actos citados en lo recuperado, con atribución de personas correcta en 23 de 24 casos.

Por su parte, la Tabla 2 resume los puntajes de la validación humana que complementa esas métricas: los quince ítems fijos cubren los tres tipos de consulta del protocolo sobre secciones diversas del corpus, con respuestas verificables en uno o dos actos; dos de los evaluadores realizaron además consultas propias de su actividad.

Tabla 2. Validación con evaluadores: media $\pm$ desvío estándar por dimensión sobre los ítems fijos (cinco jueces). Los segmentos indican ± 1 DE.

Dimensión	Media $\pm$ DE
Utilidad	4,59 $\pm$ 0,92
Fidelidad	4,82 $\pm$ 0,56
Complejidad	4,57 $\pm$ 0,74
Claridad	4,80 $\pm$ 0,57



Las cuatro dimensiones superan 4,5 sobre 5. Fidelidad (4,82 $\pm$ 0,56) y claridad (4,80 $\pm$ 0,57) concentran los puntajes más altos y los menores desvíos, en línea con la verificación determinista de citas: las observaciones de los jueces refieren principalmente omisiones de detalle —un identificador no citado, el órgano firmante ausente— antes que afirmaciones contrarias a las fuentes. La

mayor dispersión corresponde a utilidad ($\pm 0,92$) y se concentra en los ítems en los que el corpus no contiene normativa específica para la consulta y el sistema respondió con material adyacente en lugar de declarar la ausencia; los jueces coincidieron en penalizar ese comportamiento, que constituye la principal línea de mejora observada.

Los tiempos de respuesta se midieron sobre el despliegue en una máquina virtual (4 vCPU, sin GPU), separados por etapa (Tabla 3). En las consultas de un turno, la recuperación —codificación de la consulta, búsqueda híbrida y anclaje— ocupa una fracción menor del tiempo total, dominado por la generación con el modelo externo; en las conversacionales, la reescritura suma una llamada al modelo y la recuperación alcanza cerca de la mitad del total.

Tabla 3. Tiempos de respuesta en segundos sobre las 219 consultas del protocolo: mediana (p95).

Consulta	Recuperación	Respuesta completa
Un turno ($n = 161$)	0,82 (1,20)	5,44 (9,45)
Conversacional, 2.º turno ($n = 58$)	2,13 (3,20)	4,16 (6,24)

Si bien la evaluación se realizó sobre la instalación de la UNLu —donde se cuenta con mayor conocimiento del corpus y acceso a los evaluadores—, la portabilidad técnica del pipeline se verificó contra el portal SUDOCU de la Universidad Nacional de San Luis. Se generó una personalización con los datos de esa universidad —identidad institucional, esquema de colores y URL de su portal— que permitió construir una instancia completa desde el propio panel de administración: el proceso de ingesta comprobó la compatibilidad de la instalación con tres solicitudes de solo lectura, descubrió automáticamente sus 15 carpetas públicas —72.816 actos declarados— y descargó de forma escalonada, para no saturar el servidor de origen, los 26.282 documentos de tres carpetas, incluido el Rectorado, la mayor del portal.

5. Conclusiones y trabajo futuro

Se presentó un sistema real de consulta conversacional de normativa universitaria, en operación sobre el corpus completo de la UNLu y con una segunda instancia sobre el portal de la UNSL. Los aportes: un estado de diálogo cuya *procedencia* determina el peso de cada valor en la recuperación —visible, editable y con su efecto medido por ablación—; una recuperación híbrida con anclaje de identificadores normativos que resuelve por completo la consulta por identificador y encabeza la temática; una arquitectura de ingesta, indexación y operación íntegramente autoadministrada, portable entre instalaciones de SUDOCU a costo de configuración —un diseño pensado para el sistema universitario

argentino—; y correspondencia verificada entre las citas y el material recuperado, con enlace de cada cita al documento oficial publicado.

Entre las limitaciones: la evaluación se apoya en consultas sintéticas sin revisión manual, con relevancia referida al acto de origen de cada consulta; la evidencia sobre la UNSL se limita a la portabilidad técnica del pipeline; y la validación con evaluadores —cinco personas, en su mayoría vinculadas a la institución— tiene carácter exploratorio y no constituye una medida de aceptación transferible a otras instalaciones.

Como trabajo futuro se propone, por un lado, la incorporación del sistema al sistema universitario argentino —despliegue en nuevas instalaciones de SUDOUCU y evaluación con jueces de cada institución— y, por el otro, completar el corpus de la UNLu con el digesto heredado del sistema anterior, de 116.895 documentos y otra fisonomía. A ello se suma la explotación de los vínculos textuales entre actos que el sistema ya extrae —las menciones explícitas de modificación, derogación o referencia que un acto hace de otro— hacia una representación estructurada del corpus normativo, que habilitaría consultas hoy fuera de alcance: la cronología normativa de un tema, la trayectoria de una norma a través de sus modificaciones y la navegación por relaciones entre actos.

Agradecimientos. Esta investigación se desarrolló en el marco de un proyecto apoyado por la iniciativa IPAC (Cómputo de Alto Desempeño Argentina), convocatoria PCI-146 (Proyectos de Cómputo Intensivo). Los autores agradecen al Sistema Nacional de Cómputo de Alto Desempeño (SNCAD), a la infraestructura Clementina XXI, a la SICyT y su equipo técnico por el apoyo en la ejecución de los experimentos; al CIDETIC de la UNLu por la infraestructura para el despliegue institucional del sistema; y a quienes participaron como evaluadores en la validación.

Referencias

1. Anand, A., V, V., Setty, V., Anand, A.: Context aware query rewriting for text rankers using LLM. arXiv preprint arXiv:2308.16753 (2023)
2. Budzianowski, P., Wen, T.H., Tseng, B.H., Casanueva, I., Ultes, S., Ramadan, O., Gašić, M.: MultiWOZ – A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In: Proc. 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5016–5026 (2018)
3. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 2318–2335 (2024)
4. Cormack, G.V., Clarke, C.L.A., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proc. SIGIR. pp. 758–759 (2009)
5. Crompton, H., Burke, D.: Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education* 20, 22 (2023)

6. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The Faiss library. arXiv preprint arXiv:2401.08281 (2024)
7. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)
8. George, B., Wooden, O.: Managing the strategic transformation of higher education through artificial intelligence. *Administrative Sciences* 13(9), 196 (2023)
9. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20(4), 422–446 (2002)
10. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 6769–6781 (2020)
11. Kemper, S., Cui, J., Dicarantonio, K., Lin, K., Tang, D., Korikov, A., Sanner, S.: Retrieval-augmented conversational recommendation with prompt-based semi-structured natural language state tracking. In: *Proc. SIGIR* (2024)
12. Koenemann, J., Belkin, N.J.: A case for interaction: A study of interactive information retrieval behavior and effectiveness. In: *Proc. CHI*. pp. 205–212 (1996)
13. Kulesza, T., Burnett, M., Wong, W.K., Stumpf, S.: Principles of explanatory debugging to personalize interactive machine learning. In: *Proc. IUI* (2015)
14. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Proc. NeurIPS* (2020)
15. Li, T.J.J., Chen, J., Xia, H., Mitchell, T.M., Myers, B.A.: Multi-modal repairs of conversational breakdowns in task-oriented dialogs. In: *Proc. UIST*. pp. 1094–1107 (2020)
16. Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C.D., Ho, D.E.: Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies* 22(2), 216–242 (2025)
17. Manku, G., Lee-Thorp, J., Kanagal, B., Ainslie, J., Feng, J., Pearson, Z., Anjorin, E., Gandhe, S., Eckstein, I., Rosswog, J., Sanghai, S., Pohl, M., Adams, L., Sivakumar, D.: ShopTalk: A system for conversational faceted search. In: *SIGIR Workshop on eCommerce* (2022)
18. Pang, W., Wei, Z.: Exploring the transformative potential of artificial intelligence in university academic management. *Education Journal* 8(3), 97–103 (2025)
19. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* 3(4), 333–389 (2009)
20. Shuster, K., Poff, S., Chen, M., Kiela, D., Weston, J.: Retrieval augmentation reduces hallucination in conversation. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. pp. 3784–3803 (2021)
21. Thulke, D., Daheim, N., Dugast, C., Ney, H.: Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. In: *DSTC9 Workshop at AAIL-21* (2021)
22. Voorhees, E.M.: The TREC-8 question answering track report. In: *Proc. Eighth Text REtrieval Conference (TREC-8)* (1999)
23. Williams, J., Raux, A., Ramachandran, D., Black, A.: The dialog state tracking challenge. In: *Proc. SIGDIAL Conference* (2013)
24. Young, S., Gašić, M., Thomson, B., Williams, J.D.: POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE* 101(5), 1160–1179 (2013)