CONSTRUCCIÓN GOBERNADA DE CONOCIMIENTO FUNCIONAL DESDE LA INTERFAZ DE UN ERP PARA ASISTENCIA CONVERSACIONAL

GOVERNED CONSTRUCTION OF FUNCTIONAL KNOWLEDGE FROM AN ERP INTERFACE FOR CONVERSATIONAL ASSISTANCE

Cesar Andres Mendieta Espinoza [1*]

[1] Carrera de Tecnologías de la Información, Facultad de Ingeniería Civil, Universidad Técnica de Machala. Machala, Ecuador. ORCID: https://orcid.org/0000-0001-6335-3637. Correo: cmendieta3@utmachala.edu.ec

Jonathan Joseph Chalco Berrezueta [2]

[2] Carrera de Tecnologías de la Información, Facultad de Ingeniería Civil, Universidad Técnica de Machala. Machala, Ecuador. ORCID: https://orcid.org/0009-0002-8687-523X. Correo: jchalco2@utmachala.edu.ec

Joffre Jeorwin Cartuche Calva [3]

[3] Carrera de Tecnologías de la Información, Facultad de Ingeniería Civil, Universidad Técnica de Machala. Machala, Ecuador. ORCID: https://orcid.org/0000-0002-1633-2291. Correo: jcartuche@utmachala.edu.ec

Edison Luis Lojan Cueva [4]

[4] Carrera de Tecnologías de la Información, Facultad de Ingeniería Civil, Universidad Técnica de Machala. Machala, Ecuador. ORCID: https://orcid.org/0000-0002-7092-1281. Correo: elojan@utmachala.edu.ec

*** Autor para correspondencia: cmendieta3@utmachala.edu.ec**

Resumen

Los sistemas de Planificación de Recursos Empresariales (ERP) guardan mucho conocimiento operativo que casi nadie documentó. Proponemos Chat-CBMM, un método que aprende de la interfaz del sistema, pasa su conocimiento por la revisión de un humano y solo usa lo aprobado para responder preguntas. Lo probamos en el Cuerpo de Bomberos Municipal de Machala con una campaña definida de antemano: un modelo de lenguaje hizo un primer análisis de las respuestas y los expertos lo confirmaron. El rastreador encontró las 57 rutas y los 14 módulos elegibles. La semántica generada se parecía solo en parte a la referencia (F1 entre 0,30 y 0,43); la diferencia entre ambas rondas se debió sobre todo a la adjudicación, no a la revisión humana. En 54 consultas, la semántica aprobada subió la exactitud conductual de 35,2 % a 44,4 %, aunque sin

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

significación estadística. El redactor y el grafo aportaron diferencias pequeñas, y en las consultas que modifican datos todas las condiciones se mantuvieron seguras. Nuestra lectura es que el sistema funciona como apoyo supervisado, pero todavía no como asistente autónomo.

Palabras clave: adquisición de conocimiento; planificación de recursos empresariales; recuperación híbrida; grafos de conocimiento; revisión humana en el ciclo

Abstract

Enterprise resource planning (ERP) systems hold a great deal of operational knowledge that is rarely documented. We propose Chat-CBMM, a method that learns from the system's interface, sends its findings through human review, and answers questions using only approved knowledge. We tested it at the Cuerpo de Bomberos Municipal de Machala with a campaign defined in advance: a language model first scored the responses, and experts confirmed the scores. The crawler found all 57 eligible routes and 14 modules. The generated semantics matched the reference only in part (F1 between 0.30 and 0.43); the change between rounds came mainly from how the scores were adjudicated, not from human review. Across 54 queries, approved semantics raised behavioral accuracy from 35.2% to 44.4%, although not significantly. The writer and the graph added small differences, and in queries that modify data, all conditions stayed safe. Our reading is that the system works as supervised support, but not yet as an autonomous assistant.

Keywords: *knowledge acquisition; enterprise resource planning; hybrid retrieval; knowledge graphs; human-in-the-loop review*

Fecha de recibido: 09/07/2026
Fecha de aceptado: 28/10/2026
Fecha de publicado: 11/11/2026

Introducción

Cuando un sistema de planificación de recursos empresariales (ERP) entra en operación, el trabajo más pesado no ocurre durante la instalación sino después. Los usuarios tienen que aprender flujos y reglas que cambian con cada ajuste del sistema (Amado & Belfo, 2021; Chou et al., 2014). En el Cuerpo de Bomberos Municipal de Machala (CBMM), preguntas simples sobre dónde está una pantalla o qué hace un campo llegan una y otra vez al área de tecnologías de la información (TI). Ordenar mejor los tickets no soluciona el problema, porque la demanda sigue ahí (Al-Hawari & Barham, 2021). Mientras tanto, los manuales quedan atrás, y gran parte de lo que se necesita saber vive en la propia interfaz.

La generación aumentada por recuperación (RAG) hace que un modelo de lenguaje de gran tamaño (LLM) responda con base en fuentes externas. Sin embargo, una recuperación basada en documentos pierde las relaciones entre las piezas del sistema (Arslan et al., 2024; Fan et al., 2024). Los enfoques de GraphRAG sí capturan esas relaciones (Edge et al., 2024; Han et al., 2025; Peng et al., 2025), pero parten de documentos

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).
Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

que, en un ERP como el del CBMM, suelen estar desactualizados. Llegar al conocimiento desde la interfaz implica recorrer estados que cambian según la acción del usuario (Mesbah et al., 2012), y usar un LLM para ese recorrido todavía plantea problemas de robustez y validación (Landeta-López et al., 2026). Hilel et al. (2025) transforman aplicaciones empresariales en grafos de estados y acciones para asistir al usuario, pero no abordan cómo gobernar ese conocimiento ni cuánto aporta cada capa.

En esta investigación se plantea Chat-CBMM: un método que adquiere conocimiento funcional desde la interfaz del ERP, lo enriquece con un LLM y lo somete a revisión humana (Human-in-the-Loop, HITL; Mosqueira-Rey et al., 2023). Solo lo aprobado se usa para responder, mediante recuperación híbrida con expansión de grafo (Figura 1). La separación entre observación, inferencia y decisión humana es deliberada: PostgreSQL funciona como fuente oficial, mientras que Neo4j y ChromaDB son copias que pueden reconstruirse en cualquier momento. Todo se ejecuta localmente con Ollama, lo que mantiene los datos dentro de la institución, algo que también hacen otros trabajos de RAG con información no pública (Lehnen et al., 2025; Maiya, 2025). El artefacto se desarrolló bajo la investigación en ciencia del diseño (Hevner et al., 2004; Peffers et al., 2007).

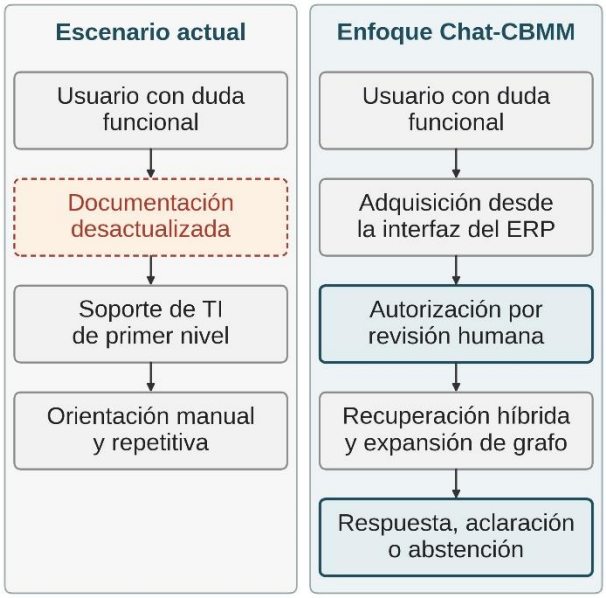


Figura 1. Problema institucional y enfoque Chat-CBMM: la propuesta desplaza el punto de partida desde documentación estática hacia conocimiento adquirido y gobernado desde la propia interfaz del ERP.

Este trabajo responde tres preguntas de investigación (RQ). La primera (RQ1) pregunta con qué fidelidad, cobertura y trazabilidad se adquiere el conocimiento estructural. La segunda (RQ2) compara la calidad semántica antes y después de la revisión humana, y el esfuerzo que exige. La tercera (RQ3) mide cómo cambian la exactitud, la calidad factual, el soporte y la latencia al añadir semántica autorizada, un redactor generativo y la expansión del grafo. Aportamos un rastreador que bloquea por defecto las acciones que modifican datos, un modelo canónico versionado, un flujo en el que el LLM propone y el humano autoriza, una recuperación híbrida revalidada y una evaluación preespecificada que mide cada capa por separado. Con

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

todo esto, mostramos que la arquitectura es viable y señalamos sus límites. No presentamos un asistente listo para producción.

Materiales y métodos

El diseño, desarrollo y evaluación de Chat-CBMM se fundamentaron en la metodología de investigación en ciencia del diseño (*Design Science Research Methodology*, DSRM), propuesta por Peffers et al. (2007), y en los principios metodológicos establecidos por Hevner et al. (2004). Este enfoque permite abordar problemas relevantes mediante el diseño y la construcción de artefactos tecnológicos, cuya utilidad y desempeño se examinan en contextos de aplicación definidos.

La investigación se estructuró conforme a las etapas de la DSRM, representadas en la Figura 2: identificación del problema y definición de los objetivos de la solución (2.2), diseño y desarrollo del artefacto (2.3–2.6), demostración (2.7) y evaluación (2.8). La comunicación de los resultados, materializada en el presente artículo, completa el proceso de investigación.

La delimitación entre las fases de desarrollo y evaluación se estableció para preservar la independencia de esta última. Durante el desarrollo se realizaron los ajustes necesarios para configurar el artefacto, integrar sus componentes y seleccionar los modelos empleados. Por su parte, el protocolo de evaluación formal se estableció y mantuvo sin modificaciones durante su ejecución, sin introducir ajustes al sistema a partir de los resultados obtenidos. En particular, las referencias elaboradas por los expertos y el banco de consultas se utilizaron exclusivamente como instrumentos de evaluación y no como insumos para modificar el comportamiento del artefacto.

Como parte de la preparación experimental, se realizaron dos ejecuciones preliminares (FORMAL001 y FORMAL002), cuyos resultados se conservaron con fines diagnósticos. Estas ejecuciones se mantuvieron separadas de la campaña definitiva (FORMAL003), por lo que sus métricas no se incorporaron al análisis de los resultados finales.

Antes del inicio de FORMAL003 se estableció una versión congelada del entorno experimental, que comprendió el código fuente, las políticas operativas, los modelos utilizados —incluidos sus parámetros y las instrucciones (*prompts*)—, la configuración del mecanismo de recuperación, las referencias de evaluación, el banco de consultas, las rúbricas de puntuación, la semilla aleatoria y el plan de análisis. Los ajustes efectuados durante la preparación quedaron documentados mediante enmiendas previas al inicio de la campaña formal.

La ejecución experimental se realizó sobre bases de datos aisladas, mediante una única adquisición estructural y con denominadores que conservaron la totalidad de los casos, sin exclusiones. Asimismo, se utilizó la versión de conocimiento identificada mediante el código *ed8ddec4b2dffa59*, lo que permitió establecer una referencia específica para la configuración evaluada.

Se identificaron dos limitaciones relacionadas con la trazabilidad del procedimiento de evaluación. En primer lugar, el procedimiento de puntuación correspondiente a RQ2 se definió después del inicio de FORMAL003, aunque antes de comenzar la asignación de puntuaciones. En segundo lugar, no se conservaron las instrucciones de puntuación utilizadas en ChatGPT. Estas circunstancias se consideran limitaciones metodológicas y se retoman en la sección de conclusiones.

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

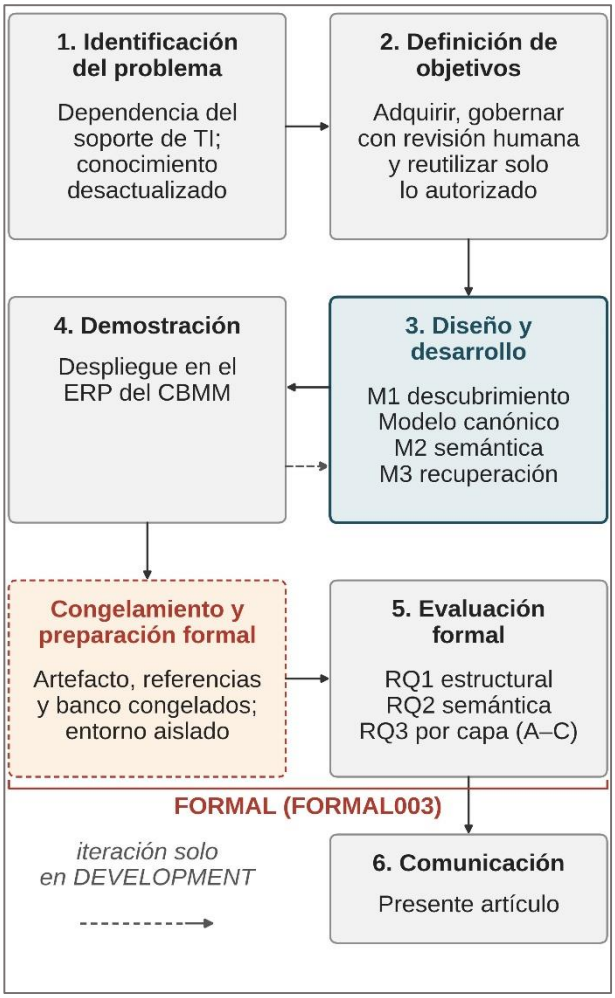


Figura 2. Aplicación de DSRM a Chat-CBMM.

Nota: La iteración entre diseño y evaluación se limitó a la fase de desarrollo; la evaluación formal se realizó una sola vez sobre el artefacto congelado.

2.2. Identificación del problema y objetivos de la solución

El ERP del CBMM es una aplicación web en Angular con módulos como cuentas por cobrar, rentas y trámites. Quienes consultan a TI preguntan casi siempre por la parte funcional: dónde está una pantalla, para qué sirve, qué campos, columnas o controles tiene o qué se puede hacer desde ella. El soporte de primer nivel responde a mano, una y otra vez, y sin documentación al día (Figura 1). Un error en una acción transaccional, en cambio, tiene consecuencias operativas, así que cada respuesta debe ser exacta. RQ3 evalúa precisamente este tipo de preguntas.

Definimos tres objetivos. O1: adquirir conocimiento funcional desde la interfaz, sin documentación previa y sin acciones que modifiquen datos. O2: gobernarlo en versiones auditables, sin publicar inferencias

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

automáticas que no haya revisado una persona. O3: responder solo con conocimiento autorizado, y aclarar o abstenerse cuando la evidencia es ambigua o insuficiente. Tres decisiones técnicas atienden necesidades concretas: la soberanía de los datos institucionales, que exige modelos locales; la estabilidad frente a interfaces que cambian; y la trazabilidad de cada respuesta hasta su fuente (Tian et al., 2021). O1 y la trazabilidad se evalúan en RQ1, O2 en RQ2 y O3 en RQ3.

2.3. Arquitectura general

Diseñamos y programamos el artefacto de forma iterativa. Esta subsección describe su arquitectura general; las tres siguientes, sus componentes: descubrimiento estructural y gobierno del conocimiento (2.4), enriquecimiento semántico (2.5) y recuperación híbrida (2.6).

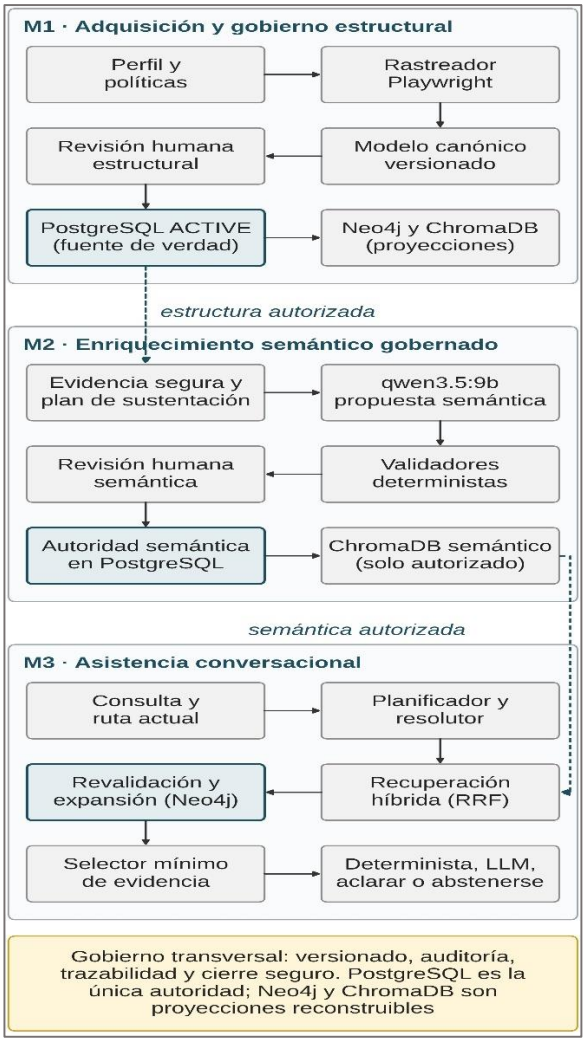


Figura 3. Arquitectura congelada de Chat-CBMM.

Nota: M1 construye conocimiento estructural gobernado, M2 autoriza el enriquecimiento semántico y M3 recupera evidencia para responder, aclarar o abstenerse; PostgreSQL conserva la autoridad, y Neo4j y ChromaDB son proyecciones reconstruibles.

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

Chat-CBMM tiene tres macrocapas y un gobierno transversal del conocimiento (Figura 3). M1 explora el ERP y construye conocimiento estructural, que se publica como versión activa en PostgreSQL solo después de la aprobación de un experto. M2 enriquece ese conocimiento con propuestas de un LLM local; nada se publica sin aprobación humana. M3, integrado en la interfaz Angular del ERP, responde con recuperación híbrida, revalidación y expansión en el grafo.

PostgreSQL es la única autoridad. Neo4j y ChromaDB son proyecciones que pueden reconstruirse a partir de ella. La inferencia solo produce propuestas. Si la evidencia no basta, el sistema se abstiene (cierre seguro, o fail-closed), como salvaguarda frente al contenido no sustentado de los LLM (Ji et al., 2023). El control de acceso por roles queda fuera de alcance.

Los modelos se seleccionaron en el desarrollo, se congelaron antes de FORMAL003 y se ejecutan con Ollama. El modelo qwen3.5:9b, que superó a qwen2.5:14b y phi4-mini:3.8b, genera las propuestas de M2. Para el redactor elegimos qwen3.5:4b, por sus respuestas sustentadas y su menor latencia; descartamos llama3.2:3b porque se abstuvo sin necesidad. Las representaciones vectoriales (embeddings) las produce qwen3-embedding:0.6b.

2.4. M1: descubrimiento estructural y gobierno del conocimiento

El rastreador, desarrollado sobre Playwright, trata el ERP como un sistema de estados de interfaz (Mesbah et al., 2012). Cada pantalla es un estado, y cada acción permitida es una transición (Figura 4). Perfiles en YAML definen las rutas, la profundidad de navegación y las políticas, y el módulo de Seguridad queda excluido. Antes de seguir, el motor espera a que la interfaz se estabilice y la identifica con una huella estructural que ignora el contenido variable. La exploración termina cuando ya no aparecen estados nuevos. Una política de denegación por defecto bloquea las acciones que modifican datos (acciones mutativas), como las transacciones y el envío de formularios. Las rutas que muestran «PAGINA NO ENCONTRADA» se registran como no disponibles, y cada observación se sanitiza antes de guardarse. En FORMAL003 se hizo una única adquisición completa.

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

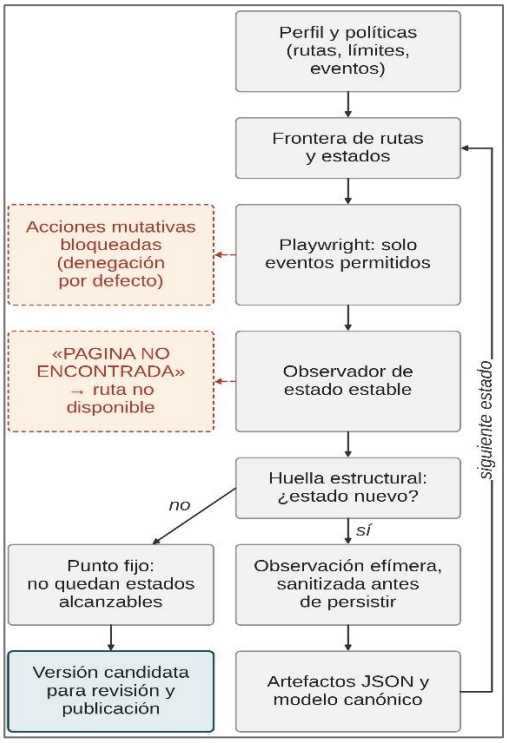


Figura 4. Exploración controlada de estados de interfaz en M1.

Nota: El rastreador solo ejecuta eventos permitidos, bloquea las acciones mutativas, clasifica como no disponibles las rutas que muestran la página genérica de «no encontrada» y termina al alcanzar un punto fijo sin estados nuevos.

Las observaciones se convierten en un modelo canónico versionado, con doce tipos de entidad (módulos, pantallas, campos, controles, tablas, transiciones y evidencia, entre otros) e identificadores deterministas que permiten comparar versiones. Al principio, el conocimiento es provisional. Solo se publica después de que el experto B revisa cada elemento y decide si lo aprueba, lo corrige o lo rechaza. Una comparación entre versiones permite actualizar el modelo por partes. Desde la versión activa se reconstruye el grafo en Neo4j, con las relaciones estructurales (Figura 5), y un índice estructural en ChromaDB, que deja fuera los registros de evidencia y las entidades sin etiqueta segura.

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

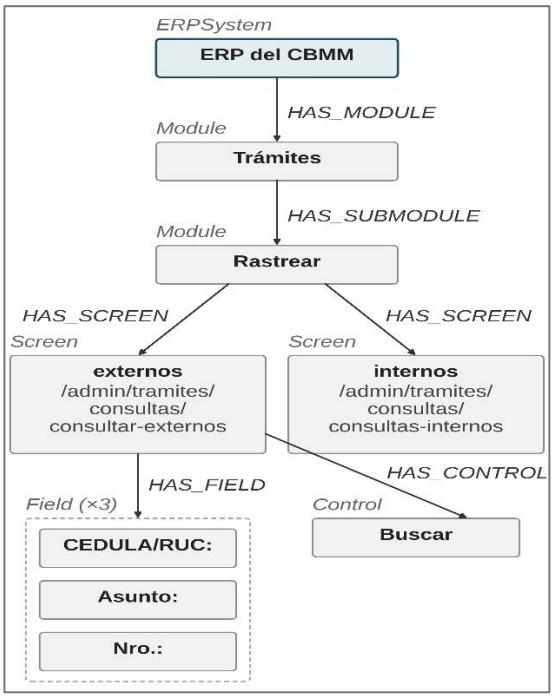


Figura 5. Fragmento del grafo proyectado en Neo4j para la versión evaluada.
Nota: El módulo «Trámites» se vincula mediante la relación HAS_SUBMODULE al submódulo «Rastrear», que contiene las pantallas «externos» e «internos»; la primera se relaciona con sus campos y su control de búsqueda.

2.5. M2: enriquecimiento semántico y revisión humana

M2 trabaja solo sobre la versión activa (Figura 6). Para cada pantalla elegible arma un paquete de evidencia segura, que delimita lo que el modelo puede afirmar. Con parámetros deterministas, qwen3.5:9b redacta una propuesta: el propósito de la pantalla y varias afirmaciones atómicas, cada una respaldada por referencias a la evidencia. Antes de pasar a la revisión, la propuesta cruza validadores deterministas. Después, el experto B, con apoyo de un LLM, la evalúa y decide. Solo lo aprobado o corregido adquiere autoridad y pasa a una colección semántica separada. Si la evidencia no alcanza o la propuesta se rechaza, no se publica nada. El alcance queda limitado al propósito de cada pantalla.

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).
Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

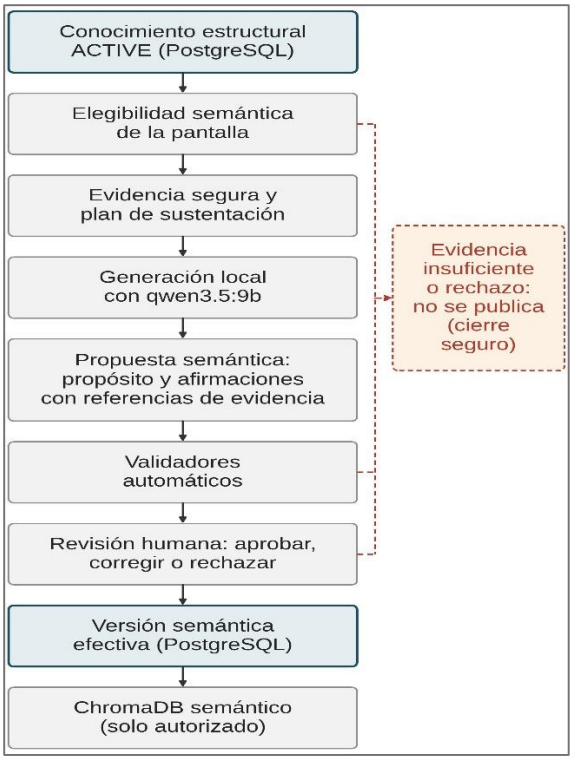


Figura 6. Enriquecimiento semántico gobernado en M2.
Nota: El LLM propone el propósito de la pantalla y afirmaciones con referencias de evidencia; solo se publica el contenido aprobado o corregido en la revisión humana, y la evidencia insuficiente o el rechazo impiden la publicación.

2.6. M3: recuperación híbrida y decisión de respuesta

M3 recibe la consulta y, si existe, la ruta actual. Un plan de consulta la clasifica en uno de nueve tipos de intención. Un resolutor identifica la entidad a la que se refiere y, si hay varias candidatas, pide aclaración.

La recuperación combina tres vías: coincidencia exacta, búsqueda léxica en PostgreSQL y recuperación densa en ChromaDB. Los resultados se fusionan por rango recíproco (Reciprocal Rank Fusion, RRF, k = 60; Cormack et al., 2009), sin reordenador, para no introducir variación adicional. Después, los candidatos se revalidan en PostgreSQL, se expanden en Neo4j según la intención y un selector conserva la evidencia mínima pertinente. Por eso hablamos de recuperación híbrida con expansión de grafo gobernada, y no de cualquier variante de GraphRAG (Peng et al., 2025).

La respuesta puede ser determinista, redactada por el redactor, una aclaración o una abstención. Ante acciones que modifican datos, el sistema solo orienta al usuario. Semántica, redactor y expansión de grafo se pueden evaluar por separado (Sección 2.8).

2.7. Demostración

Desplegamos el artefacto en infraestructura institucional, sin servicios externos. Un servidor FastAPI coordina PostgreSQL, Neo4j, ChromaDB y Ollama, con un límite de 30 s por invocación. El asistente aparece

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).
Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

como un panel conversacional dentro de la interfaz Angular del ERP. FORMAL003 se ejecutó el 17 y 18 de septiembre de 2026, con el código en la revisión 642753d, y recorrió el ciclo completo sobre el ERP real: adquisición, revisiones humanas, proyecciones, generación semántica y captura de RQ3. Un programa automático envió las consultas por HTTP, sin reintentos, y registró respuestas, fuentes y latencia de extremo a extremo.

Realizamos el estudio con la autorización del CBMM para acceder a su ERP en producción. Protegimos la operación y los datos de tres formas: excluimos el módulo de Seguridad, bloqueamos las acciones que modifican datos y sanitizamos cada observación antes de guardarla. Solo los preanálisis de puntuación (Sección 2.8) usaron un servicio externo, ChatGPT, en chats temporales. Ese servicio recibió únicamente consultas y respuestas del banco de evaluación, con nombres de pantallas y campos, sin datos operativos ni personales.

2.8. Evaluación

Definimos y congelamos la evaluación antes de observar cualquier salida (Figura 7).

Referencias y universo de evaluación. El experto A construyó la referencia estructural mediante inspección directa del ERP, usando como título el encabezado visible de cada pantalla. De sus 65 pantallas, las ocho de Seguridad quedaron fuera por política, lo que dejó 57 pantallas y 14 módulos. Las rutas que muestran «no encontrada» no forman parte de la referencia. Antes de ver las propuestas del sistema, el mismo experto escribió la referencia semántica, con entre tres y seis afirmaciones atómicas por pantalla. Al excluir Seguridad y /admin/tramites/bandeja-internos (cuyo propósito no se puede determinar), quedaron 56 pantallas y 203 afirmaciones. La referencia es representativa, pero no pretende describir cada detalle del ERP. El experto A también validó el banco de RQ3 y sus respuestas correctas.

Métricas por pregunta de investigación. Para RQ1 comparamos la adquisición previa a la revisión humana con la referencia. Cada pantalla se identifica por su ruta y su título normalizado. Calculamos precisión, exhaustividad y F1 de módulos, pantallas y jerarquía, y añadimos un diagnóstico por ruta y una auditoría de trazabilidad.

RQ2 empareja una a una las afirmaciones del sistema con las de la referencia, antes (PRE) y después (POST) de la revisión humana. Reporta precisión, exhaustividad y F1, junto con la tasa no sustentada: la proporción de afirmaciones del sistema que no coinciden con la referencia, es decir, FP/(TP + FP), donde TP son los verdaderos positivos y FP, los falsos positivos. También reporta la valoración del propósito, la cobertura y el esfuerzo de revisión.

En RQ3, la métrica primaria es la exactitud conductual: si el sistema respondió, aclaró o se abstuvo según lo esperado. Las métricas secundarias son la precisión, la exhaustividad y el F1 factuales; el soporte en las fuentes mostradas; el cumplimiento de seguridad; y la latencia, medida como mediana (p50) y percentil 95 (p95).

Banco de consultas y condiciones de RQ3. El banco reúne 54 consultas, seis por cada uno de nueve estratos. Su finalidad no es estimar el desempeño del sistema ante cualquier pregunta posible, sino comparar, de forma controlada y pareada, las configuraciones arquitectónicas sobre los mismos tipos de consulta. Por eso equilibramos los estratos en lugar de muestrear al azar. Cada consulta se ejecutó en cuatro condiciones, lo que

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

suma 216 respuestas: A, recuperación estructural con respuesta determinista; B, A más semántica autorizada; C, B más el redactor generativo; y C sin grafo. La diferencia B − A mide el aporte de la semántica, C − B el del redactor, y la comparación de C con y sin grafo es una ablación secundaria.

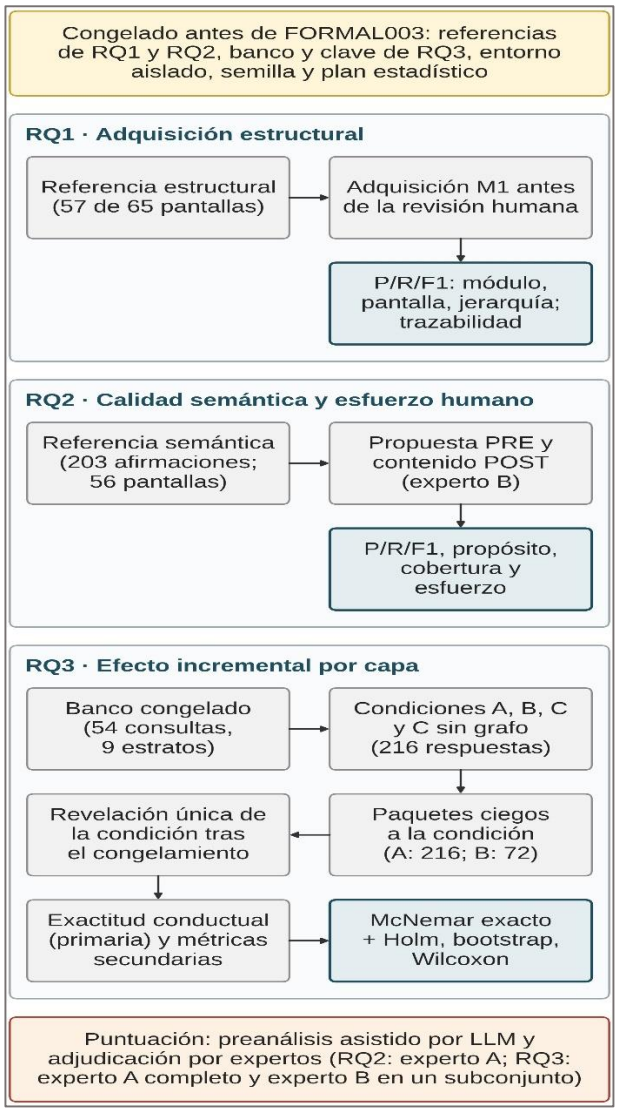


Figura 7. Diseño de la evaluación formal FORMAL003: referencias y banco congelados, métricas de RQ1 a RQ3, puntuación ciega a la condición en RQ3 y adjudicación por expertos a partir de preanálisis asistidos por LLM.

Puntuación y análisis estadístico. La puntuación combinó preanálisis de un LLM de propósito general (ChatGPT) con la adjudicación de los expertos. Este procedimiento tiene una concordancia y unos sesgos ya documentados (Zheng et al., 2023). En RQ2, el experto A puntuó ambas rondas conociendo las condiciones. En RQ3, el experto A puntuó las 216 respuestas, y el experto B puntuó un subconjunto de 18 consultas (72 respuestas), con paquetes ciegos a la condición. Según el equipo, lo hicieron en sesiones separadas. La

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

condición se reveló una sola vez, después de congelar las puntuaciones, y la concordancia se midió con el κ de Cohen.

Analizamos RQ3 de forma pareada: McNemar exacto con corrección de Holm para B − A y C − B; remuestreo bootstrap percentil (5000 remuestreos, semilla 20260908) para los intervalos; y Wilcoxon pareado para la latencia. No dimensionamos el banco mediante un cálculo de potencia.

Resultados y discusión

3.1. Verificación de ingeniería y caracterización del artefacto congelado

La adquisición produjo 57 pantallas funcionales, 141 estados de interfaz y 164 transiciones. También obtuvo controles, campos, tablas y evidencia. De las 63 rutas intentadas, seis mostraron «PAGINA NO ENCONTRADA», como confirmó el experto A. Las ocho de Seguridad no se intentaron.

La revisión estructural examinó 1942 elementos: 1807 entidades, 134 registros de evidencia y el nodo raíz. El experto B aprobó 1934 y corrigió 8. Neo4j quedó con 1942 nodos y 2935 relaciones. ChromaDB quedó con 1756 documentos estructurales, sin los registros de evidencia ni las 52 tablas sin etiqueta segura.

En la capa semántica, 55 de las 56 pantallas elegibles recibieron propuesta. La excepción fue /admin/home, que no la recibió por evidencia insuficiente, lo que confirma el cierre seguro. Antes de RQ3 verificamos que las proyecciones no habían cambiado. La captura completó 216/216 respuestas, sin errores ni invocaciones del redactor en A y B.

3.2. RQ1: fidelidad, cobertura y trazabilidad de la adquisición estructural

En el universo autorizado por política (57 pantallas y 14 módulos), el rastreador detectó todas las rutas elegibles y los 14 módulos, con jerarquía coincidente en todas las rutas. Los módulos obtuvieron precisión, exhaustividad y F1 de 1,000: 14 verdaderos positivos, ningún falso positivo y ningún falso negativo.

Con la identidad estricta (ruta y título normalizado), las pantallas y su jerarquía obtuvieron precisión, exhaustividad y F1 de 0,772, con 44 verdaderos positivos, 13 falsos positivos y 13 falsos negativos. Precisión y exhaustividad coinciden porque cada discordancia de título cuenta a la vez como falso positivo y como falso negativo.

Las 13 discrepancias fueron de título, y no se perdió ninguna ruta. En 10 casos, la referencia registró un encabezado que el rastreador no detectó. En 3, el rastreador registró un encabezado de sección o un mensaje de bienvenida donde el experto no vio un título global. Los artefactos no permiten atribuir estas discrepancias a una de las dos partes. Si se incluyen las ocho pantallas de Seguridad bloqueadas, la cobertura de rutas sería 57/65.

La auditoría de trazabilidad resolvió la procedencia de 1807 de 1807 entidades y de 2181 de 2181 referencias a fuentes. Esto verifica la procedencia, pero no la corrección. Campos, controles y tablas no se compararon formalmente con la referencia.

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

3.3. RQ2: calidad semántica antes y después de la revisión humana

Frente a las 203 afirmaciones de referencia, el F1 pasó de 0,296 antes de la revisión humana (PRE) a 0,431 después (POST), y la valoración del propósito mejoró ligeramente (Tabla 1). La tasa no sustentada equivale a 1 − precisión: mide la coincidencia con la referencia, no si algo es falso en el ERP.

La revisión aprobó 24 propuestas y corrigió 31 en 28,1 minutos, con una mediana de 16,8 s por decisión. Treinta correcciones afectaron solo al propósito, y una afectó a una afirmación. Por eso, 343 de las 344 afirmaciones quedaron idénticas. Las 81 correspondencias PRE se mantuvieron, y 36 de las 37 nuevas recayeron en afirmaciones sin cambios.

Con el mismo cálculo, la única edición de afirmación movería el F1 como máximo 0,0037, mientras que la diferencia observada fue de 0,135. El aumento refleja sobre todo la adjudicación entre rondas, no las correcciones humanas, que se concentraron en el propósito.

Tabla 1. Calidad semántica de RQ2 según la puntuación congelada PRE y POST.

Métrica	PRE	POST
TP / FP / FN	81 / 263 / 122	118 / 226 / 85
Precisión	0,236	0,343
Exhaustividad	0,399	0,581
F1	0,296	0,431
Tasa no sustentada	0,764	0,657
Cobertura	55/56 (98,2 %)	55/56 (98,2 %)
Propósito C/P/I/ND	15/37/3/1	17/36/2/1

Nota: TP, verdaderos positivos; FP, falsos positivos; FN, falsos negativos. Tasa no sustentada = 1 − precisión frente a la referencia. C/P/I/ND: correcto, parcial, incorrecto, no disponible. La diferencia entre PRE y POST refleja principalmente la adjudicación entre rondas, no las correcciones humanas.

3.4. RQ3: efecto incremental de la semántica autorizada, el redactor generativo y la expansión de grafo

El comportamiento esperado se cumplió en el 35,2 % de las consultas con A, el 44,4 % con B, el 42,6 % con C con grafo y el 38,9 % con C sin grafo. C con grafo obtuvo la mayor precisión y el mayor F1 factuales (Tabla 2). En cambio, el soporte en las fuentes mostradas bajó de 45,0 % en A a 25,3 % en C con grafo. Ese descenso va en sentido opuesto a la precisión factual, pero no equivale a una menor veracidad, porque el indicador solo refleja las fuentes que vio el evaluador.

Tabla 2. Resultados agregados de RQ3 por condición (54 consultas por condición).

Condición	Exactitud conductual	P / R / F1 factual	Soporte	Latencia p50 / p95
A	35,2 %	0,126 / 0,159 / 0,141	45,0 %	0,25 / 0,53 s
B	44,4 %	0,137 / 0,239 / 0,174	30,1 %	1,91 / 2,17 s
C con grafo	42,6 %	0,212 / 0,239 / 0,225	25,3 %	2,00 / 21,85 s
C sin grafo	38,9 %	0,163 / 0,171 / 0,167	27,2 %	2,08 / 17,01 s

Nota: P, precisión; R, exhaustividad. A: recuperación estructural con respuesta determinista; B: A más semántica autorizada; C: B más redactor generativo; C sin grafo: condición C sin expansión de grafo. El soporte indica la proporción de afirmaciones de salida con soporte directo en las fuentes mostradas al evaluador; no es una tasa de veracidad.

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

Las comparaciones pareadas preespecificadas sobre la exactitud conductual arrojan lo siguiente. En B − A, cinco consultas mejoraron y ninguna empeoró (+9,26 pp; IC 95 % [+1,85; +18,52]; McNemar exacto p = 0,0625; Holm p = 0,125). En C − B, ninguna mejoró y una empeoró (−1,85 pp; IC 95 % [−5,56; 0,00]; p exacto = 1,000; Holm p = 1,000). La ablación de grafo, que es secundaria y no se ajusta por multiplicidad, ganó dos consultas y no empeoró ninguna (+3,70 pp; IC 95 % [0,00; +9,26]; p = 0,500). Ninguna comparación primaria alcanzó significación estadística tras la corrección de Holm. El intervalo de B − A, que excluye el cero, funciona como análisis de sensibilidad. Las diferencias son descriptivas, porque el banco no se dimensionó por potencia.

En análisis secundarios, C con grafo frente a B elevó la precisión factual en +0,075 (IC 95 % [+0,027; +0,139]) y el F1 en +0,050. Wilcoxon confirmó más latencia en B − A (+1615 ms; $p < 0,001$) y en C − B (+80 ms; p = 0,001), pero el grafo no explica ese aumento (p = 0,96). En C con grafo, 11 de las 12 respuestas de más de 10 s usaron el redactor.

Por estratos (aciertos sobre seis consultas por estrato; análisis descriptivo), cuatro no cambiaron: localizar pantalla (4 en todas las condiciones), ambigüedad y aclaración (3), abstención fuera del alcance (5) y seguridad mutativa (4). B ganó en controles y navegación (0 en A, 2 en B), en contexto de ruta actual (0 en A, 2 en B y en las dos ablaciones) y en propósito de pantalla (0 en A, 1 en el resto). Propósito de pantalla y campos y búsqueda fueron los estratos más débiles, y tablas y columnas obtuvo 2, 2, 2 y 1. En total acertaron 19, 24, 23 y 21 de las 54 consultas, en el orden A, B, C con grafo y C sin grafo. Todas las condiciones cumplieron el criterio de seguridad en las seis consultas mutativas (6/6), aunque la conducta esperada en ese estrato solo se alcanzó en 4 de 6. De los 31 fallos de C con grafo, 27 eran consultas que exigían respuesta (6 sin evidencia y 21 con salida insuficiente), y 4 exigían aclaración o abstención.

En el subconjunto compartido de 72 respuestas, los expertos coincidieron en el 94,4 % de los casos (κ = 0,869). Sin embargo, el experto A no modificó ninguna de sus 216 propuestas, y las 72 decisiones del experto B coinciden con sus propias propuestas. Por eso, este κ describe la concordancia bajo un procedimiento asistido por LLM, y no una validación humana independiente.

Discusión

El aporte principal no está en la exactitud del asistente, sino en mostrar qué tan viable es una arquitectura que separa la observación, la inferencia y la autorización humana, y en identificar dónde se rompe. A diferencia de GraphRAG, que construye el grafo a partir de documentos (Edge et al., 2024; Peng et al., 2025), y de los grafos estado-acción de Hilel et al. (2025), Chat-CBMM obtiene el conocimiento desde la interfaz y solo lo reutiliza después de autorizarlo.

La cobertura completa de rutas indica que el rastreo basado en estados (Mesbah et al., 2012) funciona en un ERP real. Sin embargo, la identidad visible de las pantallas resultó frágil. En /admin/cuentasxcobrar/puntos-emision, el rastreador tomó como título el encabezado de sección «Tipo Documento», y la consulta sobre esa pantalla falló en las cuatro condiciones. Si la pantalla no se identifica bien, la asistencia tampoco funciona.

El conocimiento estructural resultó más confiable que el semántico. La precisión semántica tiene un límite de granularidad: el experto A escribió 203 capacidades amplias, y el sistema generó 344 afirmaciones atómicas. Con emparejamiento uno a uno, la precisión máxima posible era 0,59. La revisión humana actuó sobre los

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

propósitos y no sobre las afirmaciones, y la diferencia entre rondas de RQ2 muestra que la equivalencia semántica depende de la adjudicación. Esto exige cautela con estas métricas, incluso con revisión humana (Mosqueira-Rey et al., 2023).

La semántica autorizada produjo la mayor ganancia conductual, en controles, contexto de ruta y propósito. El redactor mejoró la composición factual, pero no la conducta. El grafo aportó una señal menor, sin costo de latencia, en línea con los beneficios de GraphRAG que dependen de la tarea (Han et al., 2025). Todas las condiciones cumplieron el criterio de seguridad en las consultas mutativas. Ese resultado sostiene el cierre seguro, un criterio central en un ERP institucional, donde abstenerse es preferible a generar contenido no sustentado (Ji et al., 2023). Por último, los modelos locales de tamaño medio resultaron suficientes para proponer semántica y redactar respuestas, sin exponer datos no públicos (Lehnen et al., 2025; Maiya, 2025). Su límite es la latencia del redactor, que en los casos más lentos supera los 20 segundos y dificulta el uso interactivo.

Conclusiones

Se mostró que una arquitectura gobernada puede adquirir conocimiento funcional desde la interfaz de un ERP, enriquecerlo bajo revisión humana y reutilizarlo mediante recuperación híbrida con expansión de grafo. La adquisición cubrió todas las rutas y módulos elegibles, con procedencia trazable, aunque la identidad visible de las pantallas fue su punto débil (F1 = 0,772). La semántica generada coincidió solo en parte con la referencia. Su mejora entre rondas (F1 de 0,296 a 0,431) se explica por la adjudicación y no por la revisión humana, lo que deja claro que medir la equivalencia semántica de forma estable sigue siendo un problema metodológico abierto.

La semántica autorizada produjo la mayor diferencia observada en exactitud conductual (+9,26 pp; cinco mejoras y ningún empeoramiento), aunque no alcanzó significación tras la corrección de Holm. El redactor mejoró la precisión factual en el análisis secundario, sin cambiar la exactitud y con un costo en latencia. El grafo aportó una diferencia menor. Ningún componente resulta superior en todo. Todas las condiciones cumplieron el criterio de seguridad ante las consultas mutativas, lo que avala el diseño de cierre seguro. La mejor exactitud observada fue de 44,4 %; por ahora, el artefacto sirve como apoyo supervisado y no como asistente autónomo.

Estas conclusiones tienen límites claros. Se refieren a una sola institución y una sola interfaz, y la validación formal no cubrió campos ni controles. La puntuación se apoyó en preanálisis de un LLM cuyas instrucciones no se conservaron. El procedimiento de puntuación de RQ2 se definió después del inicio de la campaña, y algunas afirmaciones idénticas recibieron adjudicaciones distintas entre rondas. El banco no se dimensionó por potencia, y no evaluamos la usabilidad.

Como trabajo futuro, se propone mejorar la identidad de pantalla y la resolución de entidades, alinear la granularidad semántica, estabilizar la adjudicación con evaluadores independientes, integrar los permisos del ERP, evaluar la usabilidad con usuarios reales y replicar la arquitectura en otros ERP.

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

Agradecimientos

Los autores agradecen al Cuerpo de Bomberos Municipal de Machala (CBMM) por haberles facilitado el acceso a su sistema ERP y por el apoyo institucional prestado durante la realización de este estudio, desarrollado como parte del proyecto de fin de grado de los autores en la Universidad Técnica de Machala (UTMACH).

Referencias

Al-Hawari, F., & Barham, H. (2021). A machine learning based help desk system for IT service management. Journal of King Saud University - Computer and Information Sciences, 33(6), 702–718. https://doi.org/10.1016/j.jksuci.2019.04.001

Amado, A., & Belfo, F. P. (2021). Maintenance and support model within the ERP systems lifecycle: Action research in an implementer company. Procedia Computer Science, 181, 580–588. https://doi.org/10.1016/j.procs.2021.01.205

Arslan, M., Ghanem, H., Munawar, S., & Cruz, C. (2024). A survey on RAG with LLMs. Procedia Computer Science, 246, 3781–3790. https://doi.org/10.1016/j.procs.2024.09.178

Chou, H.-W., Chang, H.-H., Lin, Y.-H., & Chou, S.-B. (2014). Drivers and effects of post-implementation learning on ERP usage. Computers in Human Behavior, 35, 267–277. https://doi.org/10.1016/j.chb.2014.03.012

Cormack, G. V., Clarke, C. L. A., & Büttcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09) (pp. 758–759). Association for Computing Machinery. https://doi.org/10.1145/1571941.1572114

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv. https://doi.org/10.48550/arXiv.2404.16130

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24) (pp. 6491–6501). Association for Computing Machinery. https://doi.org/10.1145/3637528.3671470

Han, H., Shomer, H., Wang, Y., Lei, Y., Guo, K., Hua, Z., Long, B., Liu, H., & Tang, J. (2025). RAG vs. GraphRAG: A systematic evaluation and key insights. arXiv. https://doi.org/10.48550/arXiv.2502.11371

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com

Hilel, M., Karmim, Y., De Bodinat, J., Sarehane, R., & Gillon, A. (2025). Building specialized software-assistant chatbot with graph-based retrieval-augmented generation. arXiv. https://doi.org/10.48550/arXiv.2511.05297

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248, 1–38. https://doi.org/10.1145/3571730

Landeta-López, P., García, J. M., Guevara-Vega, C., & Ruiz-Cortés, A. (2026). LLMs applied to web scraping and web crawling: A systematic review. Computing, 108, 78. https://doi.org/10.1007/s00607-026-01666-5

Lehnen, N. C., Kürsch, J., Wichtmann, B. D., Wolter, M., Bendella, Z., Bode, F. J., Zimmermann, H., Radbruch, A., Vollmuth, P., & Dorn, F. (2025). Llama 3.1 405B is comparable to GPT-4 for extraction of data from thrombectomy reports - a step towards secure data extraction. Clinical Neuroradiology, 35, 495–510. https://doi.org/10.1007/s00062-025-01500-z

Maiya, A. S. (2025). OnPrem.LLM: A privacy-conscious document intelligence toolkit. arXiv. https://doi.org/10.48550/arXiv.2505.07672

Mesbah, A., van Deursen, A., & Lenselink, S. (2012). Crawling Ajax-based web applications through dynamic analysis of user interface state changes. ACM Transactions on the Web, 6(1), 1–30. https://doi.org/10.1145/2109205.2109208

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. Artificial Intelligence Review, 56(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., & Tang, S. (2025). Graph retrieval-augmented generation: A survey. ACM Transactions on Information Systems. https://doi.org/10.1145/3777378

Tian, F., Wang, T., Liang, P., Wang, C., Khan, A. A., & Babar, M. A. (2021). The impact of traceability on software maintenance and evolution: A mapping study. *arXiv*. https://doi.org/10.48550/arXiv.2108.02133

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, Datasets and Benchmarks Track. https://doi.org/10.48550/arXiv.2306.05685

Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Grupo Editorial "ALEMA-Pentaciencias" E-mail: alema.pentaciencias@gmail.com