Aplicaciones prácticas de los modelos de lenguaje (LLM) en proyectos de recuperación de información

Del buscador que encuentra palabras al archivo que responde preguntas

Manuel Gómez Zotano
Director de Sistemas e Innovación · RTVE

Desde dónde os hablo

Sistemas e innovación: la parte que tiene que funcionar

Producción, no laboratorio

Sistemas reales, con usuarios reales y presupuesto público: lo que os cuente hoy está corriendo o licitado.

El archivo como sistema

MAM, catálogo, buscador, flujos documentales: la IA no llega a un vacío, llega a una arquitectura que ya existe.

Criterio, no evangelización

Tan importante es saber qué puede hacer un LLM como saber qué no le debéis pedir todavía.

Una búsqueda cualquiera en un archivo cualquiera

El usuario busca

"elecciones"

El archivo tiene

"Jornada de los comicios generales de octubre de 1982"

Resultado del buscador clásico:

0 documentos

El documento existe, está catalogado y está digitalizado. Pero el buscador encuentra palabras, no significados — y nadie escribió la palabra exacta.

Los próximos 90 minutos

3 ideas · 3 demos · 1 marco de decisión

1 · De palabras a significados (15') La búsqueda semántica: qué es un embedding y por qué cambia lo que el usuario puede preguntar. Demo en vivo.

2 · RAG: el archivo que responde (25') Vuestro fondo + un LLM que redacta citando documentos. Demo en vivo, incluida una pregunta trampa.

3 · Casos de uso reales en RTVE (20') Cinco proyectos en producción: asistente normativo, pluralismo político, Voices Unlocked, Video Identity y el grafo de conocimiento.

4 · Un marco para vuestro proyecto (10') Seis preguntas antes de escribir el pliego. Directamente aplicable en 'Plantea tu proyecto' esta tarde.

1

De las palabras a los significados

“El buscador clásico encuentra lo que se escribe igual; el semántico, lo que significa lo mismo.”

Cinco generaciones de búsqueda documental

1

Booleana

Operadores AND / OR sobre campos catalogados a mano.

años 70–80

2

Texto completo

Toda palabra del documento es indexable. Nace el 'cero resultados'.

años 90

3

Facetas y metadatos

Filtrar por fecha, programa, soporte. El estándar actual.

años 2000

4

Semántica

Se indexan significados (vectores), no solo palabras.

hoy

5

Conversacional

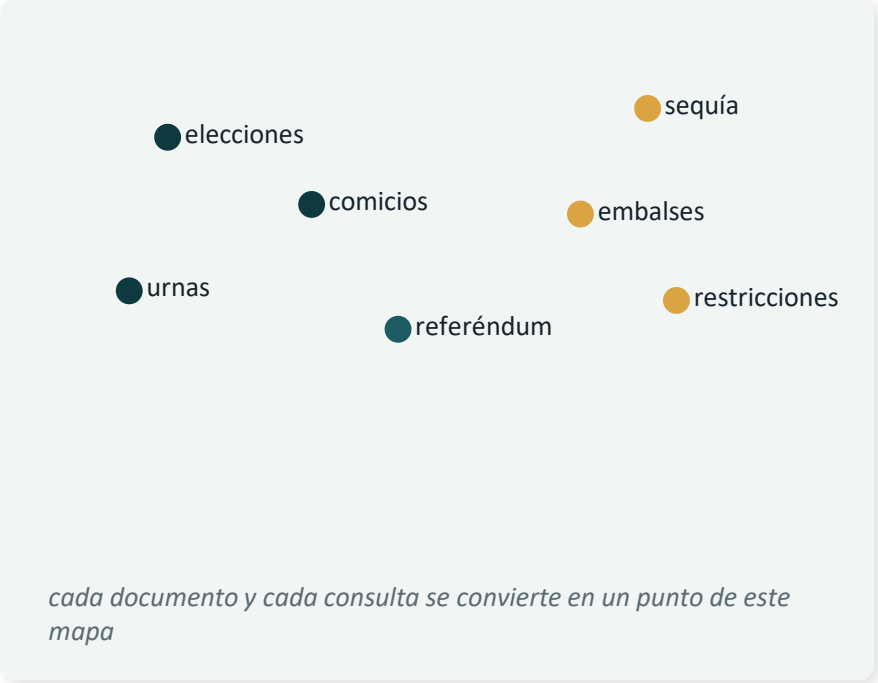
El sistema entiende la pregunta y redacta la respuesta.

hoy

Cada salto no sustituyó al anterior: lo envolvió. El buscador que necesitáis es híbrido — y los dos últimos escalones son los que trae el LLM.

El embedding: un mapa donde cerca = parecido

Toda la búsqueda semántica cabe en esta idea



cada documento y cada consulta se convierte en un punto de este mapa

Qué es Un modelo de lenguaje convierte cada texto en coordenadas. Textos que significan lo mismo caen cerca, se escriban como se escriban.

Cómo se busca La consulta se convierte en otro punto y se recuperan los documentos vecinos. 'Elecciones' encuentra los comicios del 82.

Qué NO necesita Ni recatalogar, ni diccionarios de sinónimos, ni tesauros nuevos: se indexa lo que ya tenéis, incluidas transcripciones imperfectas.

Para el desarrollador: el embedding es una función

f(entrada) → vector. Se le pasa un texto, una imagen, un audio — y devuelve números

Una sola interfaz para todo: lo que entra cambia; lo que sale, siempre, es un vector del mismo tamaño.

Determinista Misma entrada, mismo vector: se calcula una vez y se indexa. Por eso escala a millones de documentos.

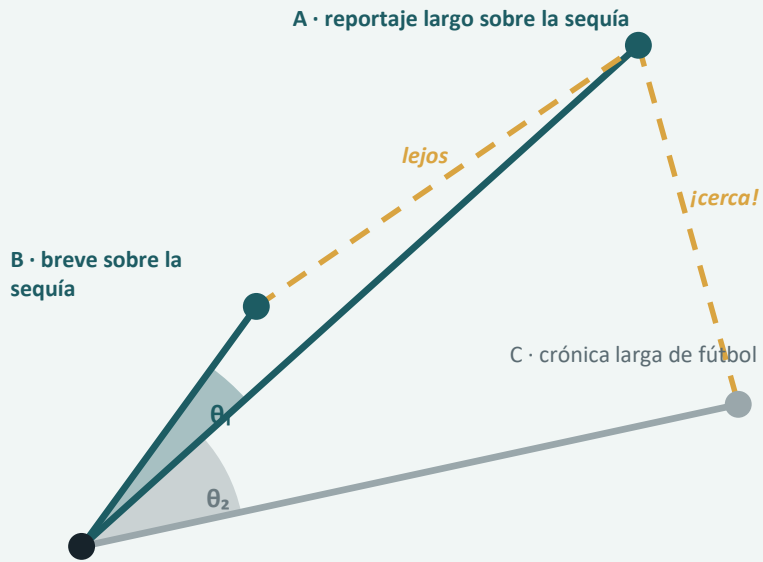
'Inyectiva' en la práctica Entradas distintas, vectores distintos (sin colisiones reales). Pero el valor es otro: lo parecido cae CERCA, no encima.

Comprime, con pérdida y de ida Salida de tamaño fijo, entre lo que entre: del vector no se recupera el documento. (Matiz de privacidad: hay inversiones parciales.)

Cada modelo, su espacio Vectores de modelos o versiones distintos NO son comparables: cambiar de modelo = reindexar. Fijadlo en el pliego.

¿Cómo se mide 'cerca'? Coseno vs. euclídea

Misma pregunta — ¿quién está más cerca de A? — dos respuestas opuestas



El coseno se lee en el ORIGEN: ϑ_1 (entre A y B) pequeño → coseno alto. ϑ_2 (entre A y C) grande → coseno bajo. El discontinuo es la euclídea: mide entre puntas.

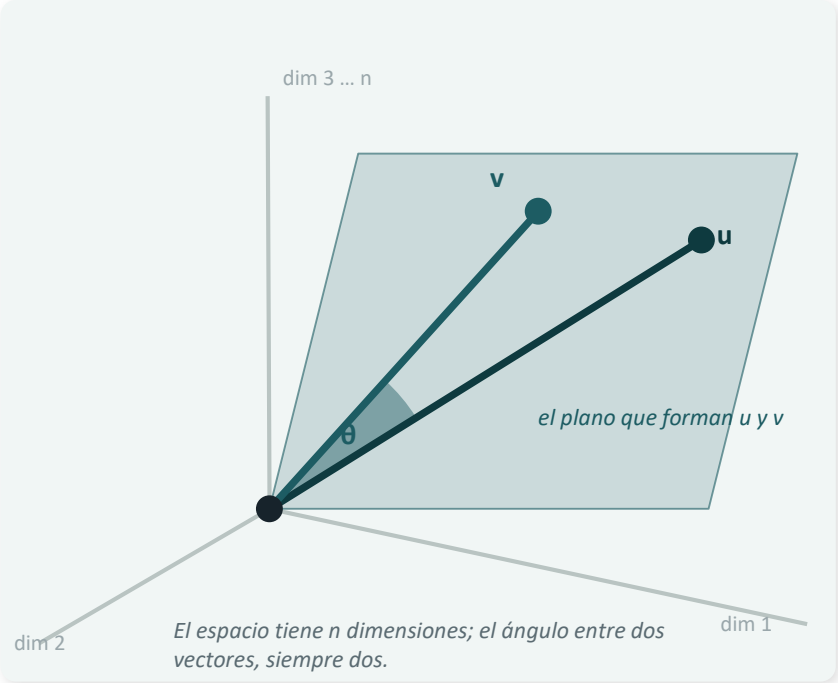
Euclídea (la regla) — ¿más cerca de A?
1º C (fútbol) · 2º B (sequía)
Mide puntas: premia medir parecido, no significar lo mismo. ¡El fútbol gana a la sequía!

Coseno (la brújula) — ¿más cerca de A?
1º B (sequía) · 2º C (fútbol)
Mide rumbos: gana quien apunta al mismo significado, mida lo que mida.

Por eso la búsqueda semántica usa coseno. (Con vectores normalizados a longitud 1, ambas ordenan igual — es lo que hace nuestra app.)

¿Y en 3 (o en 1.024) dimensiones? El plano

Dos vectores siempre forman un plano entre sí — y el ángulo vive ahí dentro



La fórmula no cambia cos θ = (u·v) / (|u|·|v|): el producto escalar dividido por las longitudes. Con 1.024 dimensiones solo hay más sumandos — no otra fórmula.

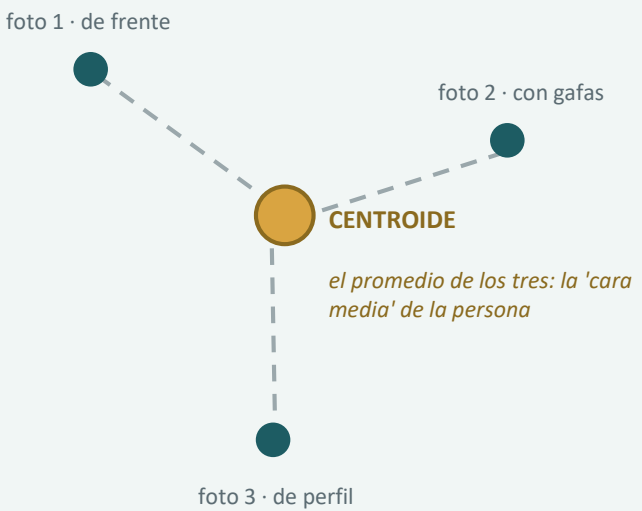
Dos vectores, un plano Tenga las dimensiones que tenga el espacio, u y v definen entre sí un plano de dos dimensiones. El ángulo θ se mide dentro de ese plano.

Por eso el dibujo anterior vale El diagrama en 2D no era una simplificación: era exactamente ese plano. La intuición visual es matemáticamente exacta.

En nuestra app los vectores tienen 384 dimensiones; en Voices Unlocked, 1.024. El ángulo, siempre entre dos.

Muchos vectores, un elemento: el centroide

Tres fotos de la misma persona → tres vectores. ¿Cuál guardo?



Si mañana llega la foto 4, el centroide se recalcula y se afina.

El problema Un mismo elemento genera varios vectores: tres fotos de Feijóo, cuatro tomas de un discurso, dos versiones de una ficha. ¿Con cuál me quedo?

La solución: promediar El centroide es la media, dimensión a dimensión, de todos los vectores: un único punto que representa al conjunto.

Por qué funciona Lo común entre ejemplos se refuerza; el ruido de cada toma (luz, ángulo, gafas) se cancela. Video Identity hace exactamente esto con las fotos de cada político.

Regla práctica: si solo puedes guardar un vector por elemento, guarda el centroide.

Qué aporta a un archivo — y una advertencia

Recuperar por concepto Sinónimos, paráfrasis, lenguaje de época ('comicios'), consultas en lenguaje natural de periodistas con prisa.

Explotar transcripciones imperfectas La semántica tolera el ruido del ASR mucho mejor que la búsqueda exacta.

Cruzar modalidades Texto que encuentra imágenes y audio: la base de los talleres 3 y 5 que ya habéis hecho.

La advertencia: híbrido siempre Para fechas, signaturas y nombres propios exactos, lo léxico sigue ganando. Lo real es léxico + vectorial + reranking. Quien os venda 'solo semántica' os está vendiendo un problema.

DEMO 1 · Léxico vs. semántico, lado a lado

Corpus: 32 registros simulados de archivo audiovisual (1968–1998)

`"sequía"`

El léxico encuentra 1 documento. El semántico añade las restricciones de agua y los embalses — sin la palabra.

`"elecciones"`

Aparecen los comicios del 82. El problema de la diapositiva 3, resuelto sin recatalogar nada.

`"posición del gobierno ante la falta de lluvia en los noventa"`

Lenguaje natural. El léxico se dispersa (trae la OTAN y una boda real); el semántico entiende la intención.

`"una multitud celebrando en la calle"`

El remate multimodal: la misma consulta ahora devuelve IMÁGENES sin metadatos ni descripciones. Texto e imagen, un solo mapa semántico.

Cierre: "Nada de esto necesitó recatalogar ni un documento — y ya mezclamos tipos de información."

2

RAG: el archivo que responde

“El conocimiento es del archivo; el LLM solo lo redacta — y lo cita.”

Anatomía de un RAG en una diapositiva

Retrieval-Augmented Generation: generación aumentada por recuperación

La respuesta se construye SOLO con vuestros documentos: el modelo no responde de memoria, responde de archivo.

“¿Y esto no es lo mismo que preguntar a ChatGPT?”

No. Y la diferencia es exactamente lo que os importa como archivo

Chatbot generalista

Responde de su memoria de entrenamiento No sabe qué hay en vuestro fondo; rellena huecos con verosimilitud.

Sin citas verificables No puede decir de qué expediente sale la respuesta.

Se queda congelado Lo nuevo que ingresáis no existe para el modelo.

Los datos viajan fuera Consultas y contenidos salen de la institución.

RAG sobre vuestro fondo

Responde desde vuestros documentos Recupera primero, redacta después: el fondo manda.

Cada afirmación, con su cita Trazabilidad documento a documento — auditable.

Siempre al día Ingesta continua: lo catalogado hoy es consultable hoy.

Desplegable en vuestra casa Local o nube privada: el contenido no sale.

Tres casos de uso que ya tienen sentido hoy

Referencia asistida

"¿Qué tenemos sobre la pantanada de Tous?" → respuesta redactada con enlaces a los expedientes y clips. El servicio de referencia, a escala.

usuarios: investigadores, periodistas

Preguntar a las horas transcritas

Miles de horas de audio y vídeo transcritas se vuelven interrogables: encontrar la declaración exacta y saltar al minuto.

usuarios: redacción, producción

Asistente interno del servicio

Normativa, cuadros de clasificación, criterios históricos de catalogación: la memoria institucional, consultable.

usuarios: el propio equipo documental

Los riesgos, sin edulcorar

Todo esto se gestiona — pero solo si se pone en el pliego y en el proyecto

Alucinaciones Las citas las reducen drásticamente, no las eliminan. Necesitáis evaluación continua con consultas de prueba y revisión humana de muestras.

Troceado que rompe contexto Un fragmento sin su jerarquía archivística puede inducir respuestas erróneas. El chunking debe respetar la descripción multinivel.

Coste por consulta Cada pregunta cuesta dinero (tokens o hardware). Dimensionar: ¿quién pregunta, cuánto, con qué modelo?

Evaluación continua Un RAG no se entrega y se olvida: se mide. Conjunto de consultas de referencia ANTES de licitar — volveremos a esto en el bloque 4.

DEMO 2 · Pregunta al archivo (RAG con citas)

LLM en local — el corpus no sale del portátil

"¿Qué material tenemos sobre la pantanada de Tous?"

Respuesta redactada citando [DOC-006] y [DOC-007]. Abrimos el contexto: la respuesta sale de ahí, no de la memoria del modelo.

"¿Qué cobertura hay de la sequía de los 90 y qué declaró el Gobierno?"

Síntesis multi-documento: combina tres fichas en una respuesta.

"¿Qué material tenemos sobre el volcán de La Palma?" · **LA TRAMPA**

No hay nada en el corpus. Un RAG bien configurado responde 'no consta' — un chatbot generalista se lo habría inventado.

Cierre: "Misma arquitectura que escala a decenas de miles de horas. Cambia el tamaño, no el concepto."

3

Casos de uso en RTVE

“Todo lo que acabáis de ver en demo, funcionando hoy — con presupuesto público y pliegos de por medio.”

Caso 1 · El RAG, en producción: asistente normativo

La demo 2 con corpus real: convenio, normas, salarios, teletrabajo, organigrama

¿Cuántos días de vacaciones tengo?

«25 días laborables de vacaciones anuales retribuidas», más días adicionales por turnicidad, territorio o disfrute fuera del periodo preferente.

[III Convenio Colectivo, Artículo 60] · Ver PDF (p. 40)

10 fuentes consultadas · 2 citadas · «orientativa: para tu caso, RR. HH.»

Lee la normativa entera Todo el corpus interno, troceado e indexado: responde citando artículo y página exacta del PDF.

Trazabilidad honesta Distingue fuentes consultadas de fuentes citadas, y avisa de sus límites: orientativo, no vinculante.

La novedad: herramientas Para el organigrama no recupera texto: llama a una herramienta que consulta el dato vivo. De RAG a RAG agéntico.

Todo lo de la demo 2 — citas, «no consta», prompt visible — más el paso siguiente: el LLM que actúa.

Caso 2 · Medir el pluralismo político

Una obligación de servicio público convertida en dato objetivo

La obligación RTVE rinde cuentas del pluralismo de su emisión ante el regulador: quién aparece, cuánto tiempo y en qué contexto.

El método Sobre transcripciones y diarización se identifica a los intervinientes y se calcula el tiempo de palabra por actor y formación política.

El resultado Minutados objetivos, continuos y trazables sobre toda la emisión — donde antes había escucha manual por muestreo.

El mismo pipeline de los talleres — transcripción + diarización — puesto al servicio de la rendición de cuentas.

Caso 3 · Voices Unlocked: activar el archivo sonoro de RNE

Una pregunta del regulador: ¿cuántos minutos dedicó RNE a promover los valores constitucionales?

Transcripción

Whisper: la voz, en texto.

Diarización

PyAnnote: quién habla y cuándo.

Segmentación

Gemma 3: temas y clasificación.

Embeddings

SpeechBrain: búsqueda vectorial.

61 h 18 min 35 s

de emisión de RNE dedicada a valores constitucionales en 2024 — calculada segmento a segmento. No una estimación: una cifra auditable.

101.447

audios indexados

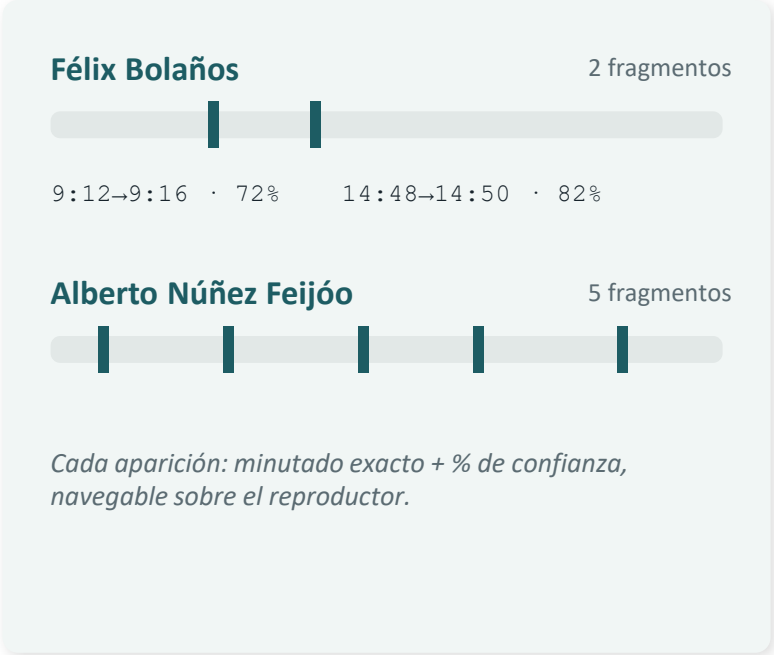
46.586

horas de emisión

El umbral de la demo, hecho ciencia: adaptativo por valor ($\mu + \alpha \cdot \sigma$, bayes-óptimo), con cada decisión registrada y reproducible ante el regulador. Toolkit de modelos abiertos, reutilizable por cualquier archivo.

Caso 4 · Video Identity: personas en el vídeo

Face ID · deepfake · contexto — el proyecto del taller 1, en producción



Tres fotos bastan Cada persona se registra con unas pocas fotos; su huella facial se recalcula como centroide al añadir más.

Buscar personas, no palabras El sistema localiza sus apariciones en los vídeos con minutado y confianza: quién sale, dónde y cuándo.

Y más que archivo La misma huella sirve para verificar: detección de deepfakes y contexto de cada aparición.

El complemento visual del pluralismo: la presencia en imagen, medible igual que el tiempo de palabra.

Caso 5 · GRAFO: el conocimiento que mejora toda la IA

El grafo de conocimiento del archivo — público en rtve.es/grafo

~3 M	26 M	~100 M	267 M
recursos digitales	entidades	relaciones	tripletas

Una capa de conocimiento común Ontología EBUCorePlus (estándar EBU): radio y TV en un único grafo. Cada persona, tema o lugar es un nodo con identidad, no una cadena de texto.

Mejora los casos anteriores Desambigua las entidades que detectan los pipelines (¿qué 'Bolaños'?), conecta apariciones, declaraciones y contenidos, y da contexto al recomendador.

Y pone barandillas al LLM GraphRAG: el RAG de la demo 2 respondiendo sobre relaciones explícitas y trazables. IA que razona, explica — y no alucina.

Hoja de ruta: buscador conversacional sobre el grafo — preguntar al archivo, con el grafo como fuente.

Cinco lecciones que nos han costado dinero aprender

1 · La transcripción condiciona todo Basura entra, basura sale: invertid en calidad aguas arriba antes que en modelos grandes aguas abajo.

2 · El documentalista valida, cura y entrena El flujo humano de revisión no es opcional: es lo que convierte metadatos automáticos en metadatos fiables.

3 · Licitar IA es distinto Métricas de calidad en el pliego (WER, precisión de entidades), no marcas ni productos. Conjunto de prueba propio para evaluar ofertas.

4 · Empezar acotado y medible Un fondo, un caso de uso, métricas definidas. 'Todo el archivo' no es un proyecto: es una promesa.

5 · El coste real no es la licencia Integración, almacenamiento vectorial, evaluación continua y personas: ahí está el 80% del coste total.

4

Marco de decisión

Seis preguntas antes de empezar

El marco para 'Plantea tu proyecto' — esta misma tarde

1 · ¿Qué pregunta no pueden hacer hoy?

El caso de uso manda, no la tecnología. Empezad por la frustración real de vuestros usuarios.

2 · ¿Qué texto tenéis o podéis generar?

Transcripciones, descripciones, OCR. Sin texto no hay LLM: a veces el proyecto empieza por transcribir.

3 · ¿Híbrido desde el día 1?

Semántica + léxica siempre. Solo-vectorial es una bandera roja en cualquier oferta.

4 · ¿Dónde viven vuestros datos?

Nube pública, privada, local: protección de datos y propiedad intelectual deciden antes que el precio.

5 · ¿Cómo mediréis que funciona?

Conjunto de consultas de prueba ANTES de licitar. Sin métrica no hay criterio de adjudicación defendible.

6 · ¿Piloto o big bang?

Siempre piloto: un fondo, un tipo de usuario, tres meses. Luego, escalar lo que funcione.

Tres peldaños de ambición (y de coste)

Se puede empezar a pilotar sin licitar: open source y herramientas comunes

a

Semántica sobre lo existente

Indexar vectorialmente lo que ya tenéis: es la demo 1.

coste y riesgo: bajos

b

RAG interno de consulta

El archivo que responde, para usuarios internos que saben detectar errores: es la demo 2.

coste: medio · riesgo: medio

c

Asistente de cara al público

Respuestas abiertas a ciudadanía: exige evaluación, tono institucional y gobernanza. No empecéis aquí.

coste y riesgo: altos

Tres ideas para llevarse

La semántica cambia la búsqueda De palabras a significados: lo ya catalogado y transcrito vale más desde el primer día.

El RAG cambia la respuesta De listas de documentos a respuestas con citas. El conocimiento sigue siendo del archivo.

La producción exige método Pliego con métricas, piloto acotado, evaluación continua y documentalistas en el circuito.

"El archivo deja de ser un lugar donde se busca y pasa a ser un interlocutor al que se pregunta. Eso no elimina el valor del documentalista: lo redefine hacia arriba."

Preguntas

Manuel Gómez Zotano · Director de Sistemas e Innovación · RTVE