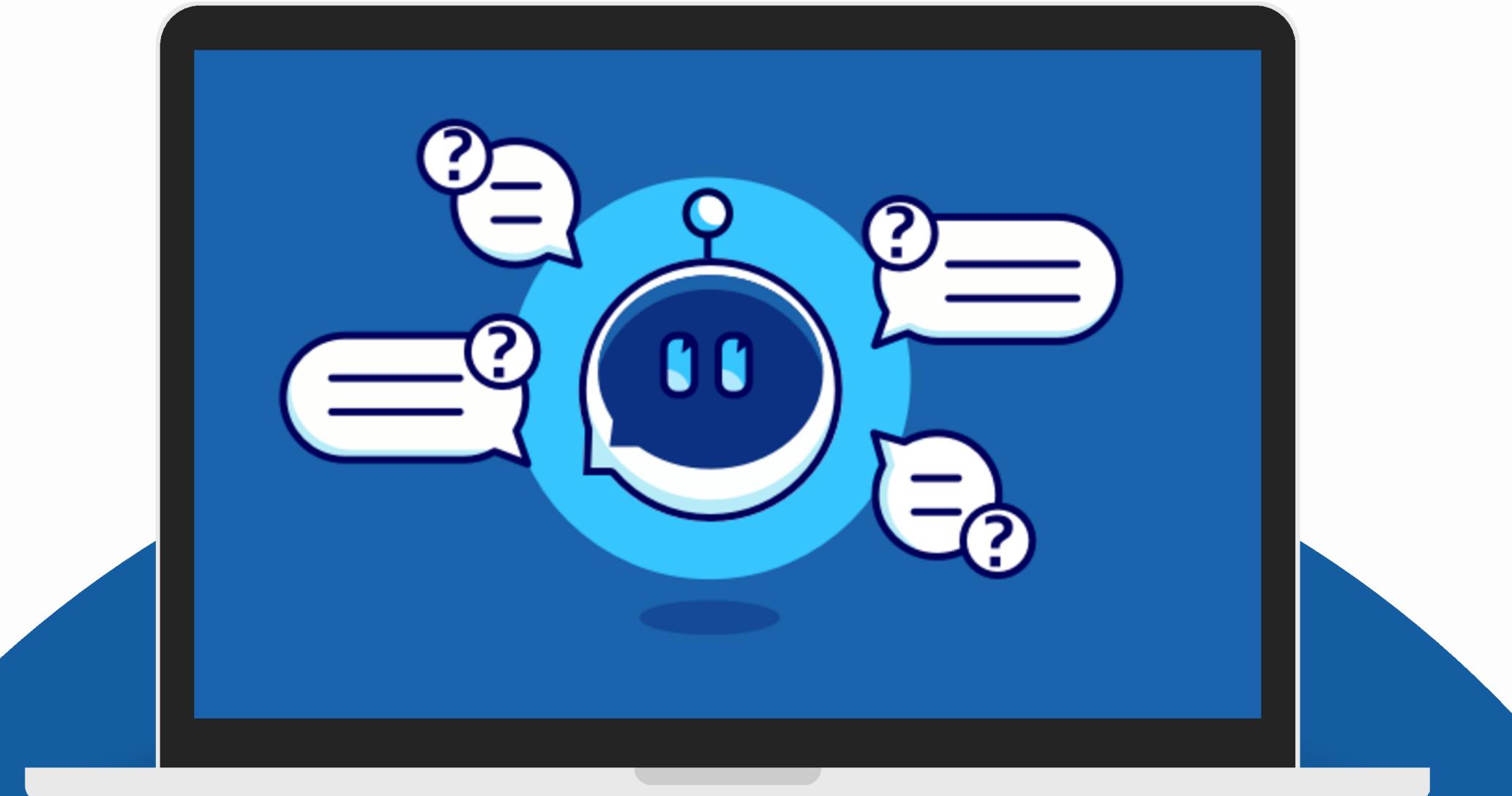




Asistente RAG

Industria EE.UU

Equipo:
Víctor Aburto
José Ignacio Muñoz
Matías Rebolledo

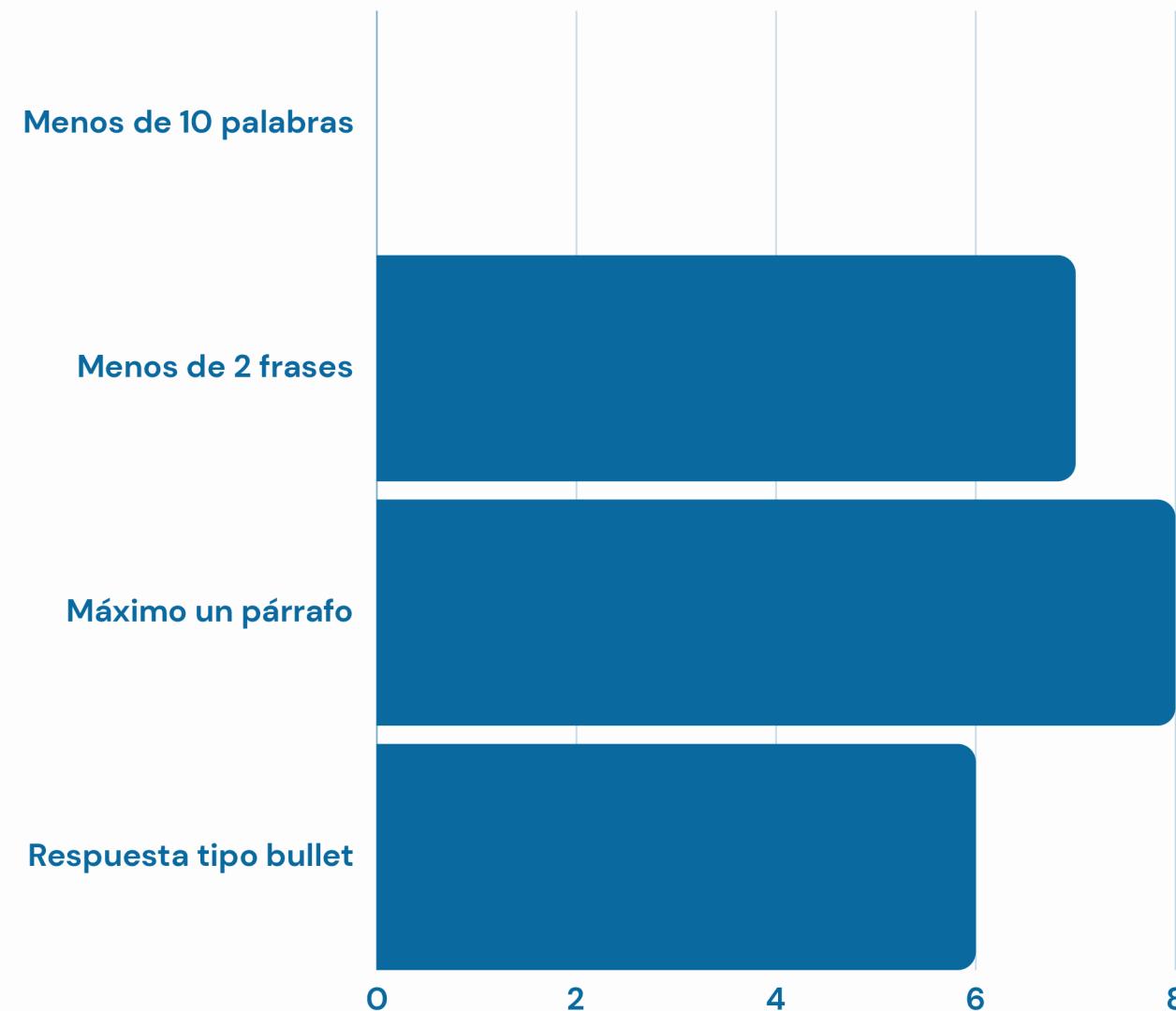


Oportunidad

Utilizar un modelo de lenguaje que pueda procesar grandes cantidades de texto y responder preguntas específicas sobre un dominio acotado.

Implementar un sistema RAG donde los usuarios puedan incluir sus documentos a una interfaz e interactuar con un chatbot que responderá sus dudas

¿Cómo deberían ser las respuestas?



Opiniones

Se destaca la importancia de la complejidad de la pregunta (propósito) y que una duración “prudente” sería entre 1 - 2 min.

57%

Se declaró usuarios intermedios de chatGPT

100%

Considera la Precisión ante rapidez

PRECISO
CONFIABLE
FIDEIDNO

SIMPLE
RÁPIDO
CLARIFICADOR
INTUITIVO

La muestra contempla 21 profesionales de distintas especialidades

Puntos clave

que deben ser considerados para implementar un sistema RAG dentro de una organización

Embedding

Representación numérica de datos en un espacio vectorial.

01

Base vectorial

Estructura embeddings para búsquedas eficientes en un espacio de características.

02

Recuperación

Busqueda de información relevante en una base vectorial para enriquecer la generación de respuestas.

03

Respuesta

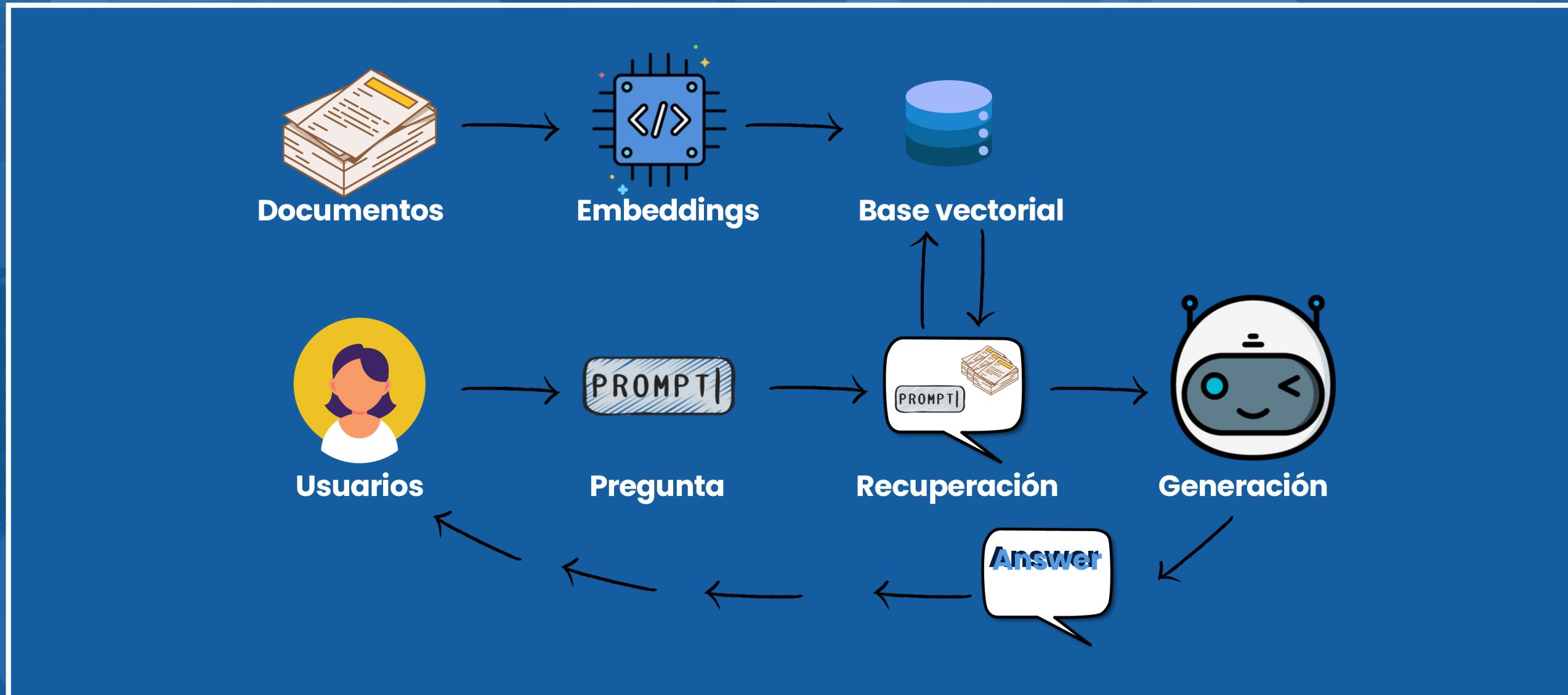
Generación de contenido final utilizando la información recuperada.

04



PIPELINE

Funcionamiento de alto nivel



¿Cómo nos hacemos cargo?

Implementando un chatbot con una interfaz que cumpla los requerimientos de negocio, pero técnicamente sea robusto.



Naive

Un modelo pequeño, más simple y con el objetivo de proporcionar respuestas rápidas



Super

Un modelo optimizado, con el objetivo de proporcionar respuestas más precisas

Esta plataforma tiene costos asociados al uso de chatGPT en el paso de generación

Super

Tiene prompt de generación que entrega un contexto amplio a la pregunta entre 400-500 palabras

Costo estimado
\$ 0.01

Modelo Fry

Este modelo corresponde a la implementación “Naive” de la solución.



Un modelo compacto con el objetivo de proporcionar una búsqueda simple y entregar respuestas rápidas.

Free

Consideraciones Técnicas

- Chunking básico:
 - Se divide un texto en segmentos en donde cada trozo se convierte en un embedding en un espacio vectorial que representa su significado.
- Búsqueda Densa :
 - Utiliza vectores generados para encontrar documentos relevantes en función de similitudes semánticas.
- Se evalúa el rendimiento , midiendo la calidad de los documentos recuperados y la precisión de las respuestas mediante distintas métricas.

Embedding

Recuperación

Evaluación



Modelo Bender

Este modelo corresponde a la implementación "Super" de la solución.



Un modelo robusto con el objetivo de proporcionar respuestas más precisas.

Costo por uso

Consideraciones Técnicas

- Chunking semántico:
 - Definición de tamaño de secuencias según buffer size.
 - Distancia coseno entre chunks.
 - Chunks según similitud.
 - Asignación de metadata a los chunks.
-
- Hybrid search :
 - Combina la búsqueda basada en palabras clave y búsqueda semántica.
-
- Se evalúa el rendimiento , midiendo la calidad de los documentos recuperados y la precisión de las respuestas mediante distintas métricas.

Demo

http://localhost:8584

The screenshot shows the MIA application interface. On the left, there's a sidebar with sections like 'Favoritos', 'Aplicaciones', 'Escritorio', 'Descargas', and 'mrebolledo'. Below that are 'Ubicaciones' (Cloud Drive, Google Drive), 'Etiquetas' (super, naïve, Rojo, Naranja, Amarillo, Verde, Azul, Morado, Gris), and 'Subir documentos' (with a 'Drag and drop files here' area and a 'Limit 200MB per file • PDF, TXT, DOC' note). At the bottom of the sidebar are 'Evaluación del modelo' and 'Evaluar modelo' buttons.

The main area has a 'Hoy' section showing a file named 'CFR-2024-vol8.pdf' (Grabación de... la(s) 22.10.04). Below it is a 'Últimos 7 días' list with several capture files. A modal window is open over the list, displaying the content of the PDF file 'CFR-2024-vol8.pdf'. The PDF is titled 'SUBCHAPTER L—REGULATIONS UNDER CERTAIN OTHER ACTS ADMINISTERED BY THE FOOD AND DRUG ADMINISTRATION' and includes sections like 'PART 1210—REGULATIONS UNDER THE FEDERAL IMPORT DRUG ACT', 'Subject A—General Provisions', '§ 1210.1 Authority', '§ 1210.2 Scope of part', and '§ 1210.3 Definitions'.

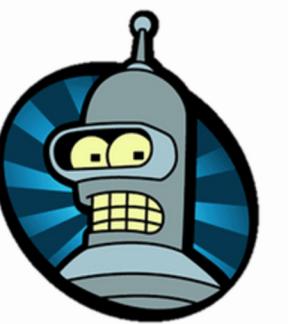
Below the PDF preview, the file information is summarized: 'CFR-2024-vol8.pdf' (Documento PDF - 343 KB), 'Información', 'Creación' (hoy, 21:38), 'Modificación' (hoy, 21:38), 'Última apertura', and 'Etiquetas' (with an 'Agregar etiquetas...' button). At the bottom of the modal are 'Mostrar opciones', 'Cancelar', and 'Abrir' buttons, with 'Abrir' being the one currently being clicked.



Métricas de resultado

En la tabla se observan las métricas asociadas a:

- Threshold: 0.3
- Temperature: 0.5
- Buffer: 2

Modelo	Context precision	Context recall*	Faithfulness	Answer relevancy
	7%	6%	81%	92%
	24%	6%	91%	95%

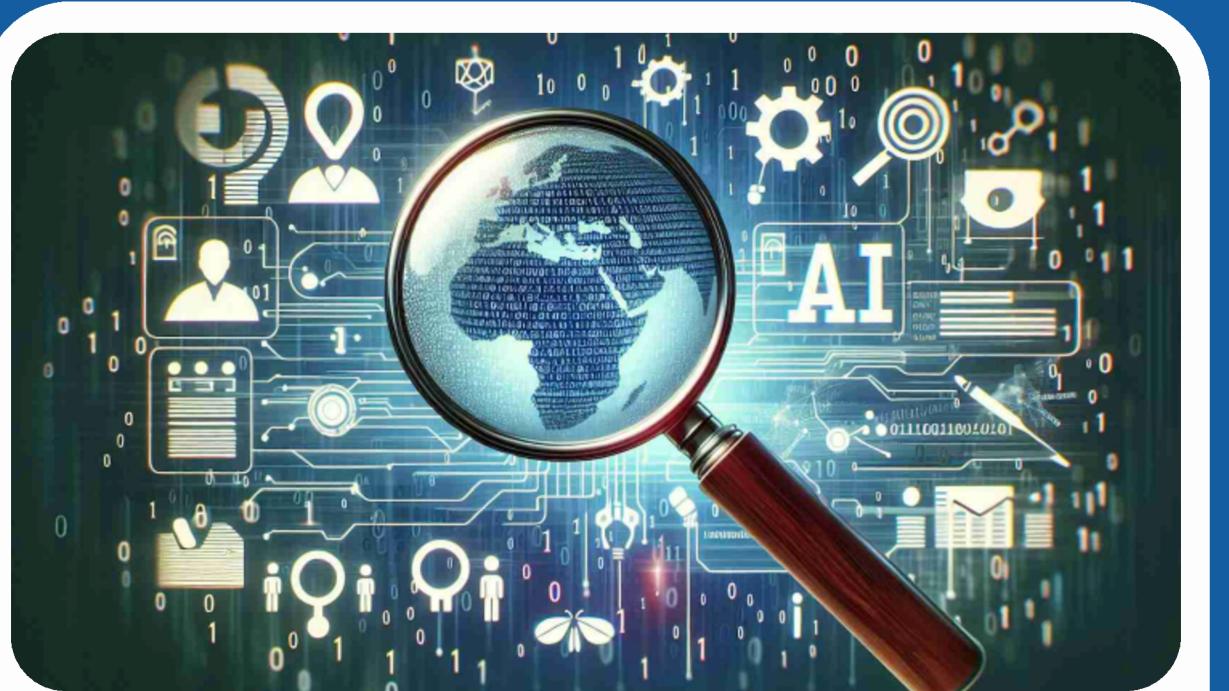


Próximos pasos



Base de vectores

Escalar la solución a una herramienta con nube



Recuperación

Ponderación de elemento densos y elementos de coincidencia exacta



Data analytics

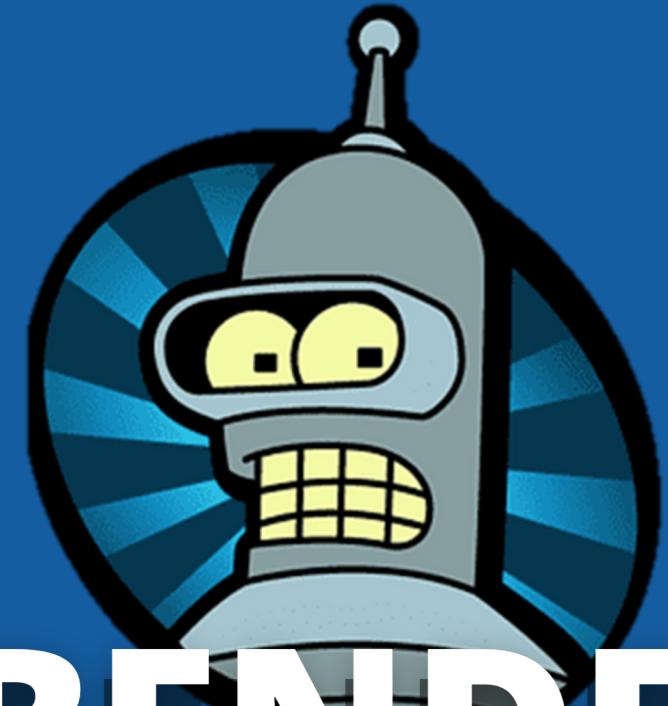
Monitoreo de las metricas de recuperación.

NEW RELEASES

20.12.2024



FRY



BENDER

by:

