

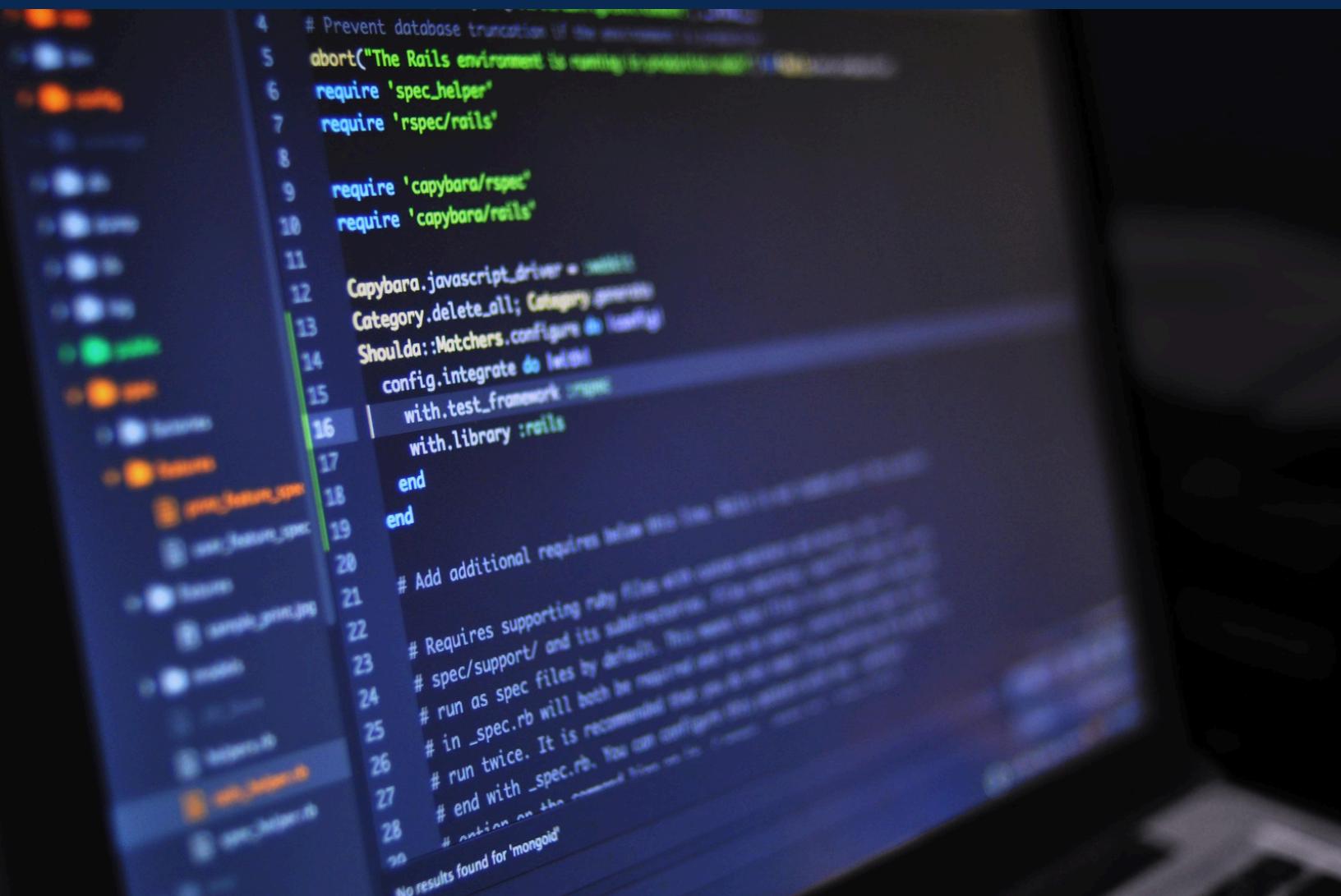


BÚSQUEDA DE RESPUESTAS EN DOCUMENTOS DIGITALIZADOS

• BRIANNA AYELEN BALAM VELASCO
• JESÚS JAVIER CAN NOH

APRENDIZAJE COMPUTACIONAL 2025

ÍNDICE DE CONTENIDOS



1. MOTIVACIÓN
2. OBJETIVO
3. PIPELINE
4. ¿QUÉ MIDE DUDU?
5. DEMO

1. MOTIVACIÓN

- Document VQA requiere razonamiento con información distribuida y no literal.
- Las métricas actuales de recuperación (p. ej., Chunk Score@k) evalúan coincidencia de texto o presencia literal.
- Estas métricas fallan cuando la respuesta es abstractiva o requiere inferencia multi-salto.
- Necesitamos una métrica que evalúe si lo recuperado permite deducir la respuesta, no solo si contiene la cadena exacta.



2 . O B J E T I V O

- Diseñar DUDU (Deductive Understanding of Document Underpinnings) , una métrica que evalúa si lo recuperado en RAG contiene evidencia suficiente para inferir la respuesta (no solo “hacer match” de texto).
- Integrarla al flujo de Document VQA con RAG textual + visual (layout) para casos abstractivos y de razonamiento.
- Validar que correlaciona mejor con humanos que métricas de coincidencia literal (Hit@k, Chunk Score@k, nDCG@k) y que diferencia sistemas aunque generen la misma cadena.



Example 1 (correct prediction)

Q: "How many rats were fed the control diet?"



GT: ["ten", "ten male rats"]
(pag. 7)

Textual retrieval

Top 3 chunks:

"8 june 18, 1975 the ability of dietary glycan to lower elevated blood cholesterol and triglycerides in rats. **ten male rats** were fed the control diet (basal + cholesterol and cholic acid) for four weeks and divided into two groups. one group (2a) continued to receive the control diet. the second group (2b) received a test diet of the same"

"second group (2b) received a test diet of the same composition as the control diet except that 15 parts of glycan replaced 15 parts of sucrose. another group (#1) of ten male rats were fed the basal diet for four weeks and then continued on the basal diet at weekly intervals, whole blood was drawn from the caudal vein for"

"fiber 2 2 20 ---- glycan --- 2 5 10 20 cholesterol + 0.2% cholic 0 1.0 1.0 1.0 1.0 1.0 males: feed consumption (g) 126 124 144 126 123 121 124 014 cholesterol consumption (g) 1.24 1.44 1.26 1.23 1.21 1.24 cholesterol excreted (g) .046 .664 .861 .526 .603 .625 .750 % excreted"

A: "ten male rats"

Visual retrieval

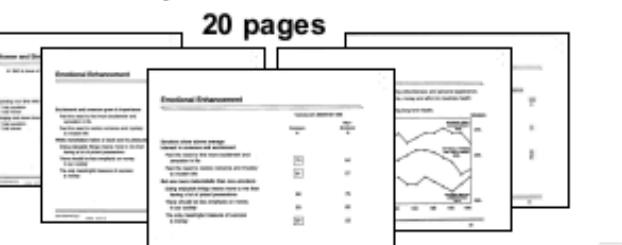
Top 5 image patches[†]:



A: "ten"

Example 2 (wrong prediction)

Q: "What percentage of non-smokers feel there should be less emphasis on money in our society?"



GT: ["82%", "82"]
(pag. 7)

Textual retrieval

Top 3 chunks:

"non-smokers doing enjoyable things means more to me than having a lot of prized possessions 68 75 **there should be less emphasis on money in our society** 80 82 the only meaningful measure of success is money 27 23 rj712/monitor/mgj/n 29 7886 9e ets"

"to pleasure doing enjoyable things means more to me than having a lot of prized possessions 67 71 73 **there should be less emphasis on money in our society** 27 31 33 the only meaningful measure of success is money 27 25 24 rj712/monitor/mgj/n 28 1886 9e ets"

"the physical self = physical appearance trends undergoing a similar shift in emphasis focus on looking good and being well groomed - over the long term - with minimum effort = signs of turning away from fashion perfectionism less attention to "latest fashions" less competitiveness more interest in comfort - physical - emotional = smokers are as involved in appearance as non-smokers rj712/monitor/mgj/n 32 8886 9e ets"

A: "31"

Visual retrieval

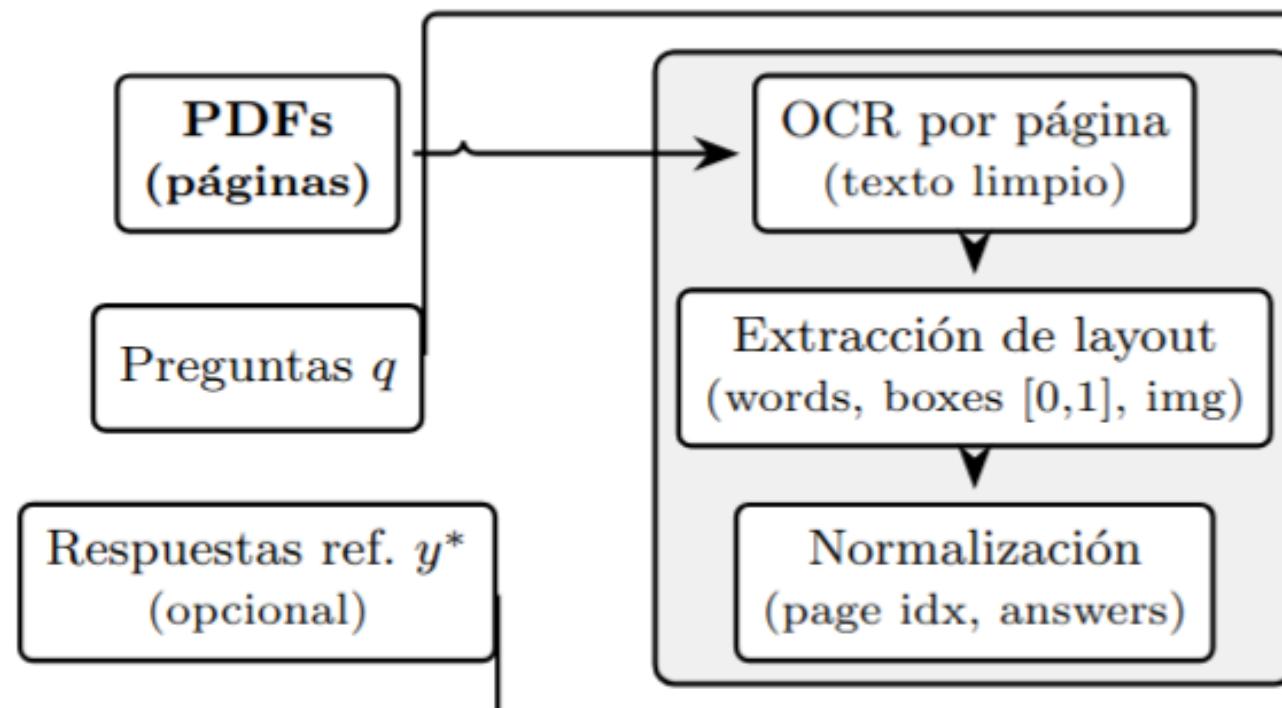
Top 5 image patches[†]:



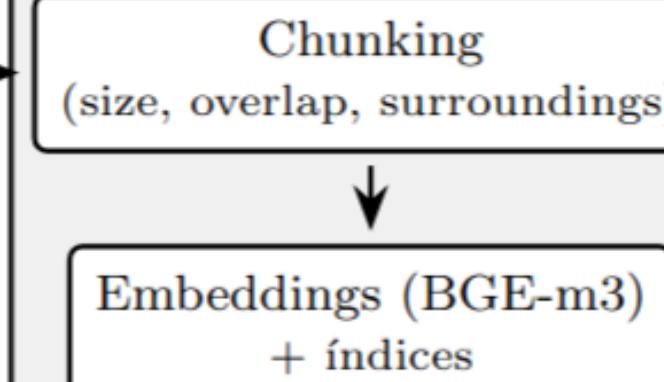
A: "80"

3. PIPELINE

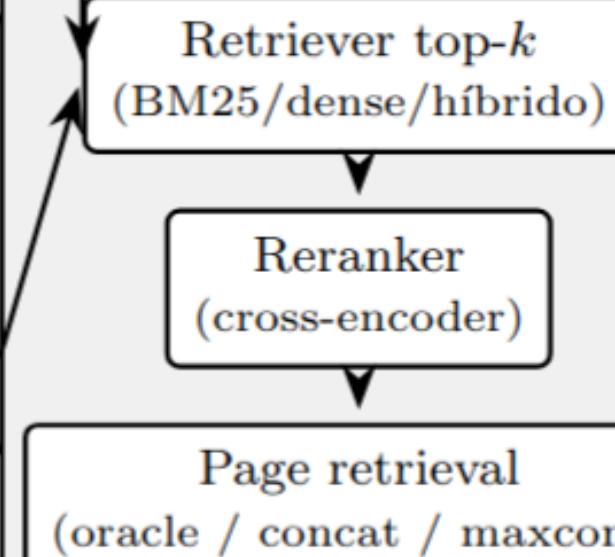
OCR & Layout



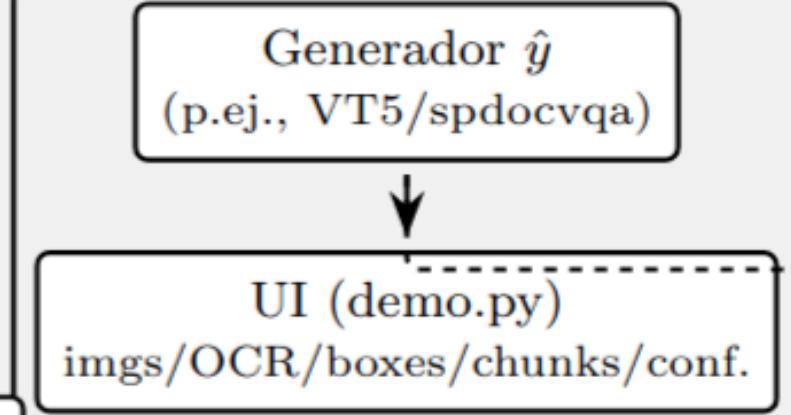
Chunking & Embeddings



Retrieval



Generación & UI



DUDU — LLM Judge

Entrada: $(q, y^*, R = \{r_1..r_k\}, \hat{y})$
Salida JSON: {suficiencia 0–5, consistencia 0–5, alucinación 0–2, justificación}

4. ¿QUÉ MIDE DUDU?



DUDU = DEDUCTIVE UNDERSTANDING OF DOCUMENT UNDERPINNINGS

Evalúa si lo recuperado por RAG contiene evidencia suficiente y alineada para inferir la respuesta de referencia (aunque no haya string-match).

- Problema: Hit@k / nDCG@k premian coincidencias literales; fallan en respuestas abstractivas/derivadas.
- Propuesta: un LLM-judge puntuá Evidencia → Razonamiento → Respuesta.

Meta: mayor correlación con humanos y mejor sensibilidad a cambios reales del retrieval (no solo del generador).

DUDU evalúa la capacidad de deducir la respuesta usando solo la evidencia recuperada y su coherencia, no la elocuencia del generador.

5 . D E M O

RAG Visual Question Answering Demo (MyPDFThreads)

Load Next Batch

Question ID

101

Load Sample

Original Information

Original Page Images

processing | 13.0/253.8s

Original OCR Text

RAG Visual Question Answering Demo (MyPDFThreads)

[Load Next Batch](#)

Question ID

101

Load Sample

Original Information

Original Page Images | novità | como te sentiste al usarlo? - que tal tu
idea navegar? - notaste problemas de rendimiento? gracias, jordan
customer success Asunto: Re: Seguimiento a tus comentarios (app
móvil) De: Reese (Cliente) Para: Jordan (Miembro de equipo) Fecha: 2
de marzo de 2025, 3:44 p. m. CST hola jordan, me gusto el diseño.
algunas acciones en notificación se sintieron algo lentas. saludos,
rees

escrito. Asunto: Re: Re: Notas de la conversación - appmovil.be.
Taylor (Líder de equipo) Para: Jordan (Miembro de equipo) Fecha: 2 de marzo de 2025, 9:49 p. m. CST hola jordan, por favor, da seguimiento para saber si el problema de rendimiento ocurre en todos los notificaciones o solo en algunos. gracias, taylor team lead el 2 de marzo de 2025, 7:59 p. m. cst, jordan (miembro de equipo) escribio

Original OCR Text

team lead
el 13 de marzo de 2025, 12:21 a. m. cst, kendall (miembro de equipo)
escribio:
bien.

Page 173:

10/12/25, 2:12 pm hilos de correo -- retroalimentacion de producto bien.

escribio:

lentitud ocurre en todos los busquedas o solo en algunos?

escribio:

rendimiento ocurre en todos los busquedas o solo en algunos.

equipo) escribio:

escribio:

seguro.

Question and Answers

Question	Actual Answers	Actual Answer Page Index	Predicted Answer	Predicted Answer Confidence	Retrieved Page Indices
[ID: 101] ¿Quién envió el último correo?	Harper (Líder de equipo) Para: Kendall (Miembro de equipo), Harper (Líder de equipo) Para: Kendall (Miembro de equipo)	0	emerson	0.16663618385791779	[[171]]

Use via API 🔮 · Built with Gradio 🎨

Question	Actual Answers	Actual Answer Page Index	Predicted Answer	Predicted Answer Confidence	Retrieved Page Indices
¿En qué fecha se envió el último correo?	13 de marzo de 2025, 5:49 a.m. CST	0	12 de marzo de 2025, 12:21 p.m.	0.32935652136802673	21

Use via API 🔮 · Built with Gradio 🎨

Question and Answers

Question	Actual Answers	Actual Answer Page Index	Predicted Answer	Predicted Answer Confidence	Retrieved Page Indices
[ID: 101] ¿Quién envió el último correo?	Harper (Líder de equipo) Para: Kendall (Miembro de equipo), Harper (Lider de equipo) Para: Kendall (Miembro de equipo)	0	emerson	0.16663618385791779	[[171]]

Use via API 🔥 · Built with Gradio 🎨

```
args = {
    "model": "RAGVT5",
    "dataset": "MyPDFThreads",
    "model_name": "RAGVT5",
    "dataset_name": "MyPDFThreads",

    "data_dir": pdf_root,
    "preprocessed_dir": "./data/cache",
    "gt_csv_path": "data/mypdfs/gt_threads_dev.csv",

    # RAG / retrieval
    "use_RAG": True,
    "embed_model": "BGE",
    "embed_weights": "BAAI/bge-m3",
    "reranker_model": "",
    "reranker_weights": "cross-encoder/ms-marco-MiniLM-L-6-v2",

    "page_retrieval": "concat", # "concat", "maxconf"
    "add_sep_token": False,
    "batch_size": 1,
    "layout_batch_size": 4,

    # chunking
    "chunk_num": 10,
    "chunk_size": 40,
    "chunk_size_tol": 0.2,
    "overlap": 20,
    "include_surroundings": 0,
    "retrieval_top_k": 20,      |
    "reranker_top_k": 8,

    "model_weights": "rubentito/vt5-base-spdocvqa",
    "lora_weights": "",

    "cache_dir": "./hf_cache",
}
```

1. MANTENER BILINGÜE (ES/EN).
2. GENERADOR (VT5) EN GPU FP16 PARA VELOCIDAD.
3. EMBEDDINGS (E5-SMALL) EN CPU (BARATO EN RAM/VRAM).
4. RERANKER EN GPU PERO CON HUELLA MÍNIMA (POCOS PARES Y MICRO-BATCH).

ESCALAMIENTO PARA MEJORA

8 GB (INTERMEDIO)

- EMBEDDINGS A GPU (EMBED_DEVICE="CUDA:0")
- RETRIEVAL_TOP_K=15-20 · RERANKER_TOP_K=8-10
- LAYOUT_BATCH_SIZE=2 · RERANKER_MAX_LENGTH=256-384
- MENOR LATENCIA Y MEJOR RANKING

12-16 GB (ALTO)

- TODO EN GPU (EMBEDDINGS + RERANKER + VT5) CON FP16/AUTOCAST
- LAYOUT_BATCH_SIZE=3-4 · RERANKER_TOP_K=10-20
- RERANKER_MAX_LENGTH≈512 · (OPCIONAL) BGE-M3 PARA MÁXIMA CALIDAD

IMPLEMENTACION CON LLMS

 Phase 1 — Gemini FAISS + Local Cross-Encoder Reranker

Retrieval con embeddings de Gemini (API) -> rerank local (cross-encoder) -> Gemini LLM. k'=20 → k=10 por defecto; ajustable con env vars.

q

Resume en 3 viñetas los principales problemas mencionados en los hilos de correo.

ClearSubmit

output

Aquí están los principales problemas mencionados:

- Problemas de rendimiento en las notificaciones.
- Notificaciones que tardan más de lo esperado.
- Lentitud en la carga de exportaciones. | Sources: Hilos de Correo — Retroalimentación de Producto.pdf p.6, Hilos de Correo — Retroalimentación de Producto.pdf p.77, Hilos de Correo — Retroalimentación de Producto.pdf p.3, Hilos de Correo — Retroalimentación de Producto.pdf p.5, Hilos de Correo — Retroalimentación de Producto.pdf p.172, Hilos de Correo — Retroalimentación de Producto.pdf p.105

Examples

[Resume en 3 viñetas los principales problemas mencionados en...](#)[¿Quién reportó problemas de rendimiento y en qué contexto? C...](#)[Summarize the main product feedback in 3 bullets, with citat...](#)[List issues related to search and under which conditions the...](#)

IMPLEMENTACION CON LLMS

FASE 1

Phase 1 – Gemini FAISS + Local Cross-Encoder Reranker

Retrieval con embeddings de Gemini (API) -> rerank local (cross-encoder) -> Gemini LLM. k=20 → k=10 por defecto; ajustable con env vars.

q

Resume en 3 viñetas los principales problemas mencionados en los hilos de correo.

Clear Submit

output

processing | 1.2/3.5s

Examples

Resume en 3 viñetas los principales problemas mencionados en... ¿Quién reportó problemas de rendimiento y en qué contexto? C... Summarize the main product feedback in 3 bullets, with citat... List issues related to search and under which conditions the...

MUCHAS GRACIAS

