PDF Retrieval at Scale
with Visual Language Models

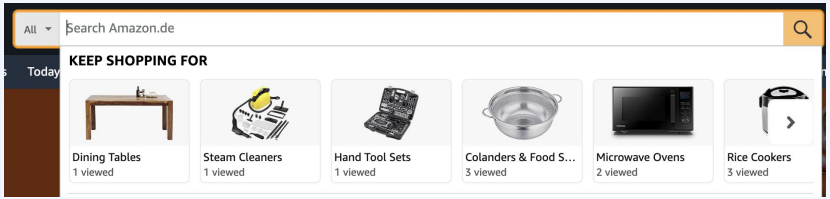
Evgeniya Sukhodolskaya,
Developer Relations,
Qdrant

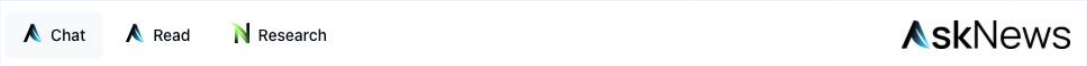
Agenda

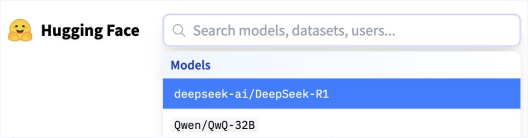
*"Discover **how(4) Visual Language Models like ColPali and ColQwen(3)** unlock new possibilities in **PDF(2) retrieval(1)** at **scale(5)**."*

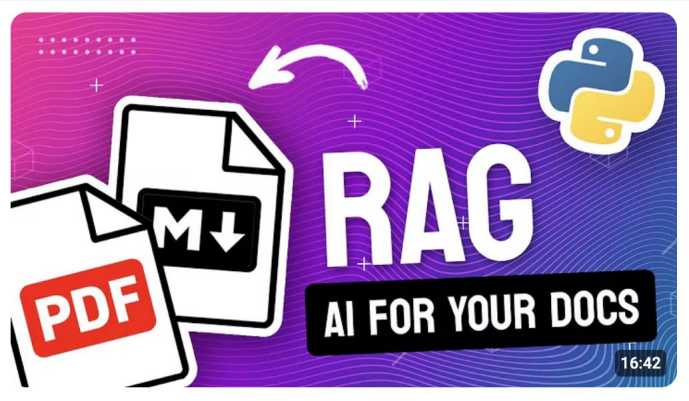
Let's Start with Retrieval

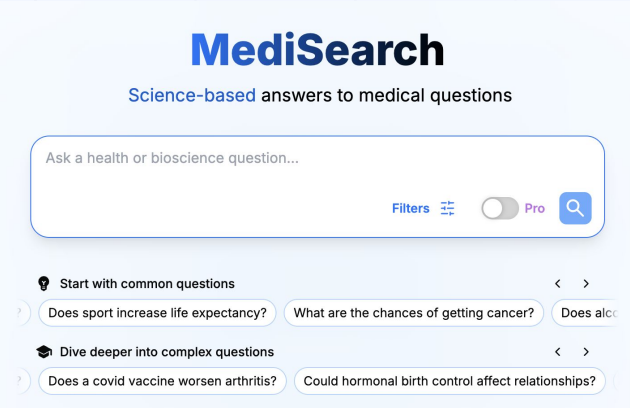
Retrieval is Everywhere



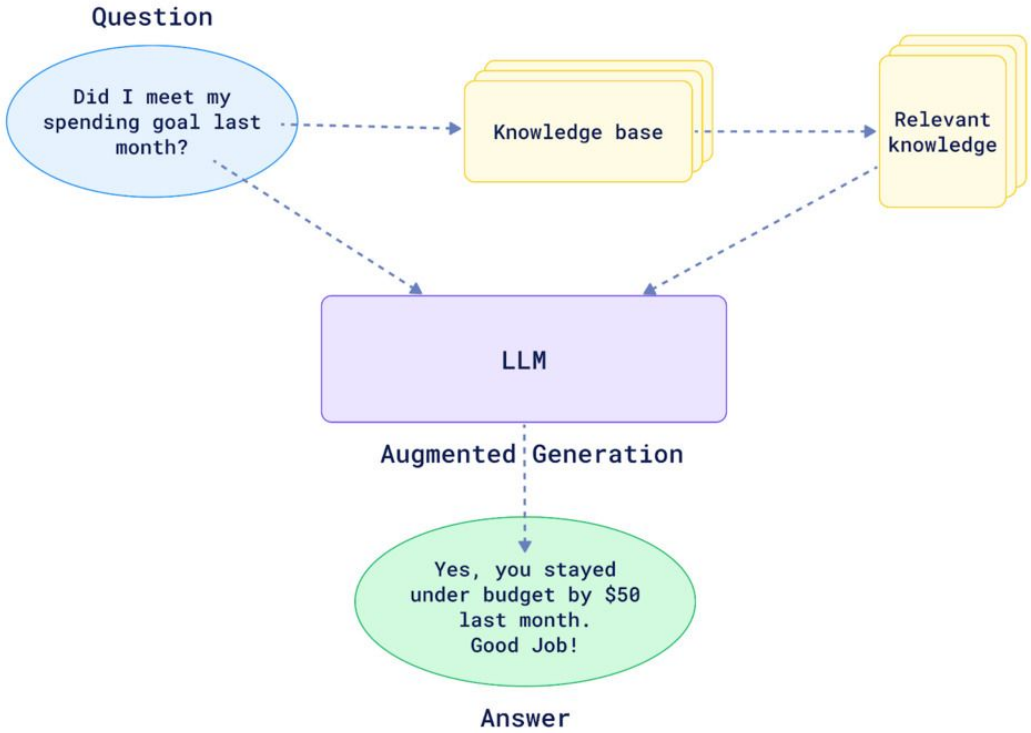








Now we have Conversational Retrieval



Why now?

*The era of **conversational retrieval** was driven by:*

A. *"Efficient Estimation of Word Representations in Vector Space"*, 2013
B. *"Attention is All You Need"*, 2017
C. *"Scaling laws for neural language models"*, 2020
D. *"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks"*, 2020
E. *Reptilians*, 600 BC

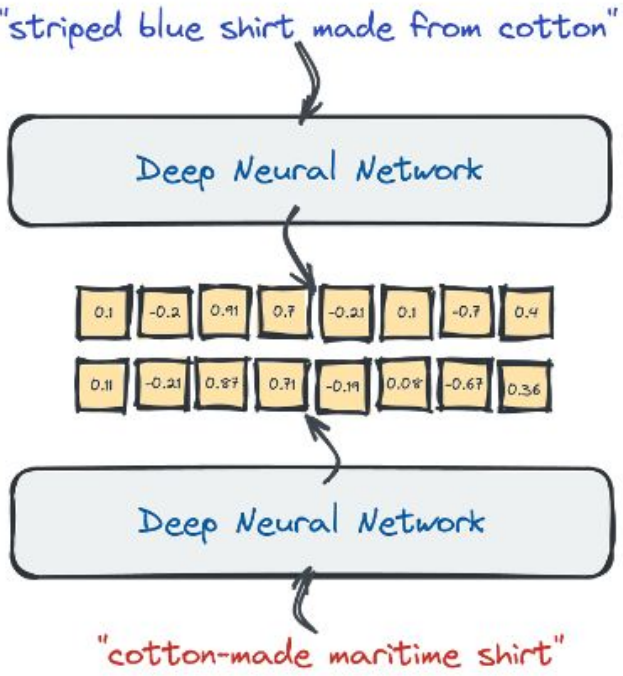
Why now?

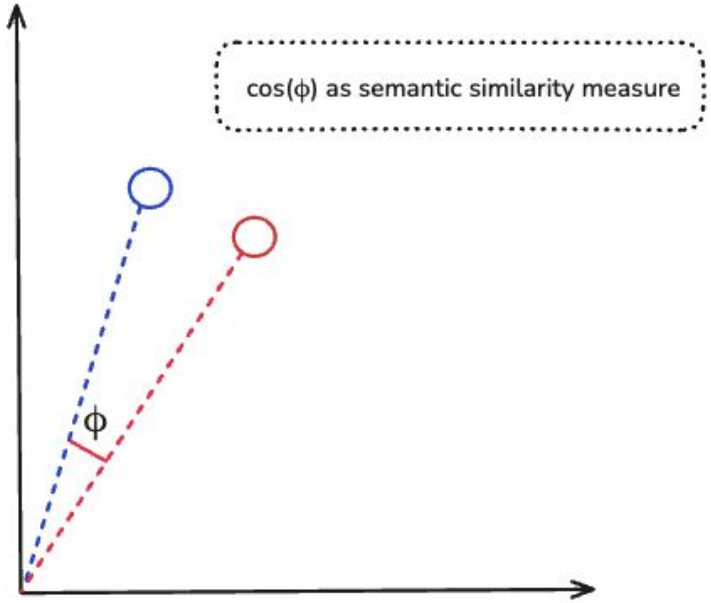
The era of **conversational retrieval** was driven by:

A. *"Efficient Estimation of Word Representations in Vector Space"*, 2013
B. *"Attention is All You Need"*, 2017
C. *"Scaling laws for neural language models"*, 2020
D. *"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks"*, 2020
E. *Reptilians*, 600 BC

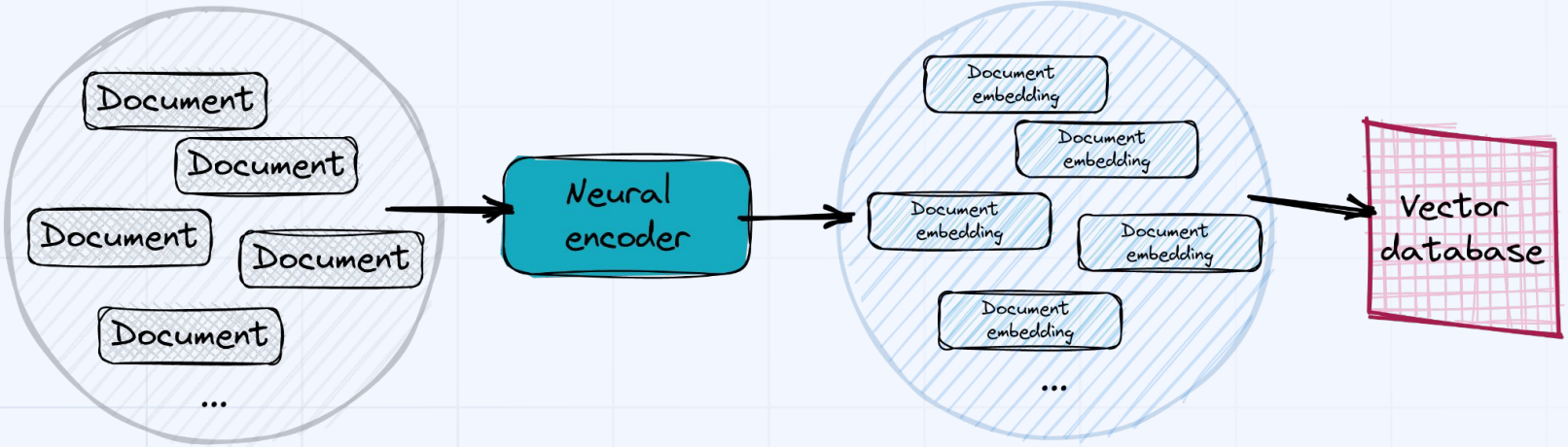
All in all, imho – by **scalable semantic search** + **"smart" generative models**.

Semantic Similarity Search

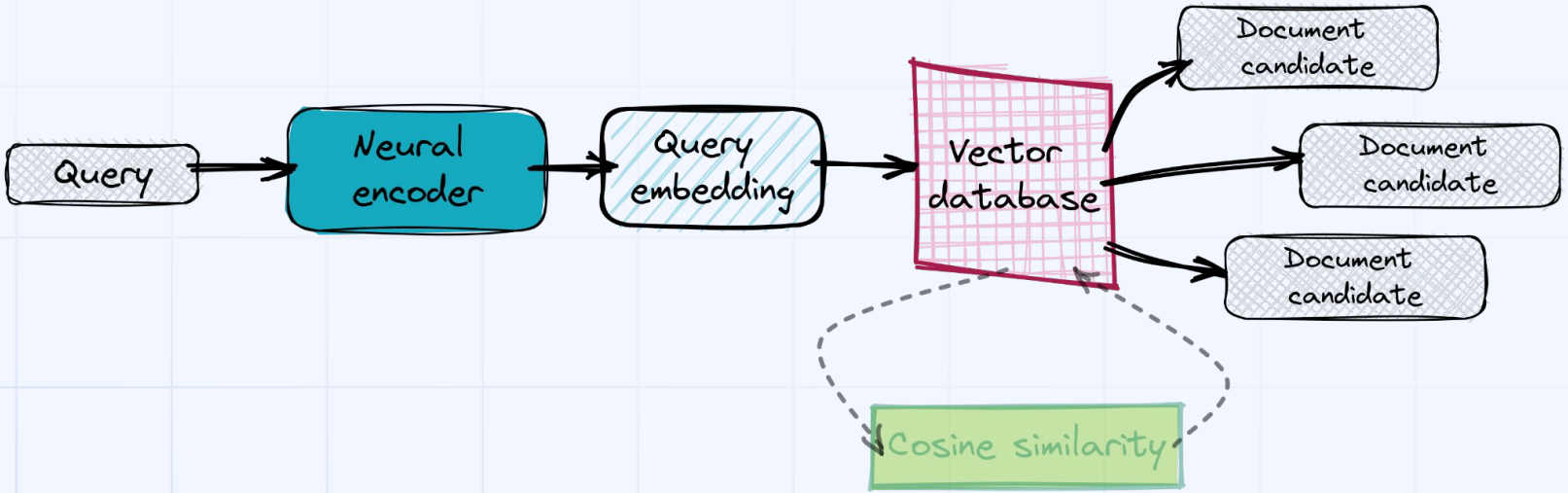




Semantic Search at Scale: Part 1 – Indexing



Semantic Search at Scale: Part 2 – Retrieval



PDF (Conversational) Retrieval

PDFs are also Everywhere

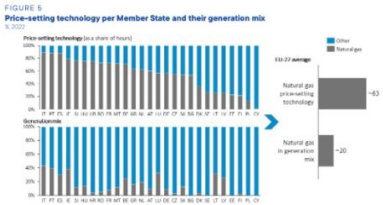
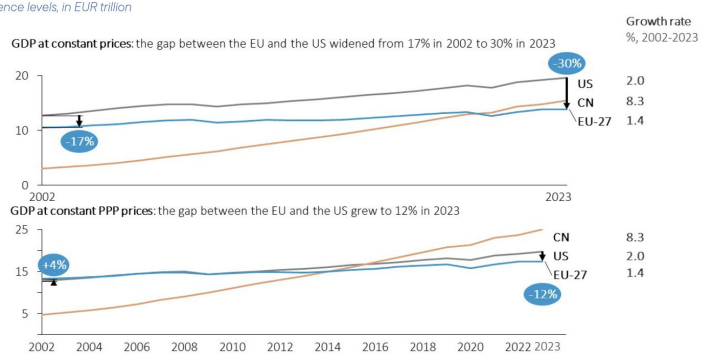
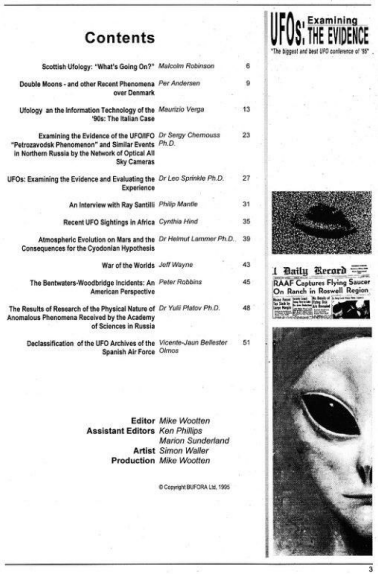


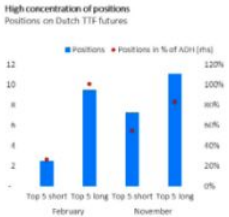
FIGURE 3
GDP evolution
2015 reference levels, in EUR trillion



Source: OECD, 2024.

02. The value of the gap in GDP in any given year is only indicative. It should not be viewed as an exact estimate as price deflators and purchasing power adjustments are imperfect. When comparing GDP developments across countries, the price deflator and exchange rate have an important effect on results. Depending on the objective of the comparison, one or the other indicator may be more relevant. GDP at current prices offers insights into market value, GDP at constant prices into volume growth, while purchasing power adjustment allows a comparison from the consumer perspective.





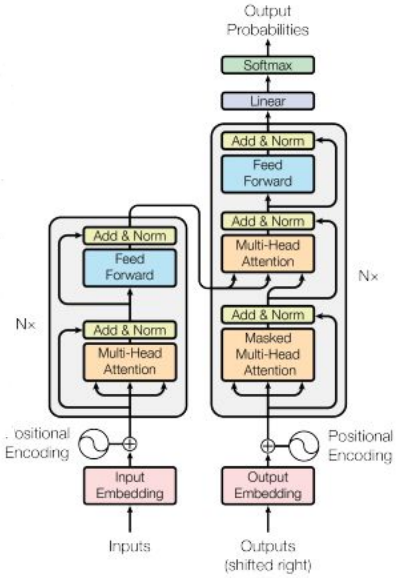


Figure 1: The Transformer - model architecture.

he Transformer follows this overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, shown in the left and right halves of Figure 1, respectively.

Possibilities with PDF Retrieval

1. Chat with research papers;
2. Find answers in business reports;
3. Quickly analyze study materials;
4. Medical/legal assistants;

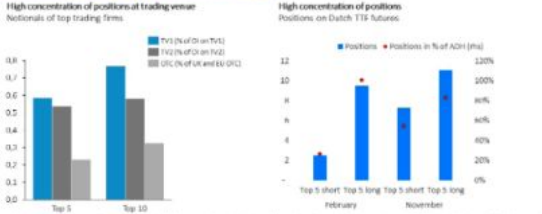
...

Search and retrieve graphs, images, tables, text, charts, and logos from large archives.

Standard PDF Retrieval: A Typical Workflow

FIGURE 4

Market concentration in EU gas derivatives markets



Note: Market share of natural gas by venue in % of reported notionals, excluding central counterparties and clearing members. The figure shows that the top-5 and top-10 EU counterparties (in terms of gross notionals) accounted for more than 50% and 60% respectively of reported notionals by EU entities on each of the two EU gas regulated markets. Data as of November 2022. OI: Open Interest. TV: Trading Venue. OTC: Over-the-counter.

Note: Absolute value of net positions in EUR billion for the top five long and short non-financial corporate counterparties and positions in % of average daily trading volume, in % rhs. The high concentration of positions indicates that if several firms with similar directional positions were to reduce their exposures, they could amplify market moves.

Sources: Trade repositories (TRs), Bank of England, ESMA.

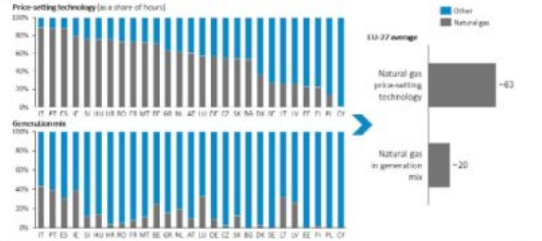
Sources: EMIR, ESMA.

Europe's market rules pass on this volatility to end users and may prevent the full benefits of decarbonising power generation from reaching them. Even as Europe reduces its dependence on natural gas and increases investment in clean energy generation, its market rules in the power sector do not fully decouple the price of renewable and nuclear energy from higher and more volatile fossil fuel prices, preventing end users from capturing the full benefits of clean energy in their bills [see Figure 5]. In 2022 at the peak of the energy crisis, natural gas was the price-setter 63% of the time, despite making up only 20% share of the EU's electricity mix. The use of long-term contract solutions – like Power Purchase Agreement (PPA) markets or Contracts for Difference (CfDs) – can help attenuate the link between the marginal price setter and the cost of energy for end users, but such solutions are underdeveloped in Europe, in turn limiting the benefits from accelerating the roll-out of renewables. In the absence of action, this decoupling problem will remain acute at least for the remainder of this decade. Even if renewable installation targets are met, it is not forecast to significantly reduce the share of hours during which fossil fuels set energy prices by 2030.

FIGURE 5

Price-setting technology per Member State and their generation mix

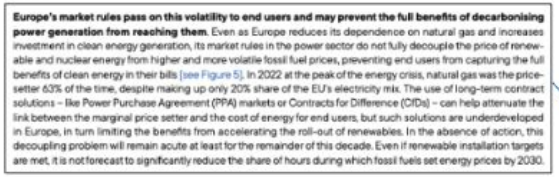
%, 2022



Source: European Commission (JRC), 2023.

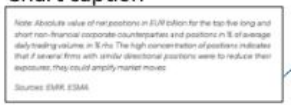
1. Optical Character Recognition (OCR)
2. Layout detection
3. Processing visual elements

Text



Europe's market rules pass on this volatility to end users and may prevent the full benefits of decarbonising power generation from reaching them. Even as Europe reduces its dependence on natural gas and increases investment in clean energy generation, its market rules in the power sector do not fully decouple the price of renewable and nuclear energy from higher and more volatile fossil fuel prices, preventing end users from capturing the full benefits of clean energy in their bills [see Figure 5]. In 2022 at the peak of the energy crisis, natural gas was the price-setter 63% of the time, despite making up only 20% share of the EU's electricity mix. The use of long-term contract solutions – like Power Purchase Agreement (PPA) markets or Contracts for Difference (CfDs) – can help attenuate the link between the marginal price setter and the cost of energy for end users, but such solutions are underdeveloped in Europe, in turn limiting the benefits from accelerating the roll-out of renewables. In the absence of action, this decoupling problem will remain acute at least for the remainder of this decade. Even if renewable installation targets are met, it is not forecast to significantly reduce the share of hours during which fossil fuels set energy prices by 2030.

Chart caption



Note: Absolute value of net positions in EUR billion for the top five long and short non-financial corporate counterparties and positions in % of average daily trading volume, in % rhs. The high concentration of positions indicates that if several firms with similar directional positions were to reduce their exposure, they could amplify market moves.

Sources: EMIR, ESMA.

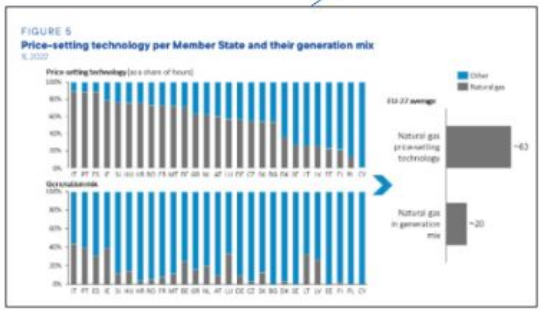
Text embedding model

1. Optical Character Recognition (OCR)
2. Layout detection
3. Processing visual elements

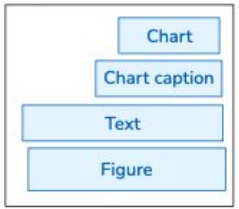
Chart

Good visual embedding model for charts(??)

Figure



Layout



Text

Europe's market rules pass on this volatility to end users and may prevent the full benefits of decarbonising power generation from reaching them. Even as Europe reduces its dependence on natural gas and increases investment in clean energy generation, its market rules in the power sector do not fully decouple the price of renewable and nuclear energy from higher and more volatile fossil fuel prices, preventing end users from capturing the full benefits of clean energy in their bills [see Figure 5]. In 2022 at the peak of the energy crisis, natural gas was the price-setter 63% of the time, despite making up only 20% share of the EU's electricity mix. The use of long-term contract solutions – like Power Purchase Agreement (PPA) markets or Contracts for Difference (CfDs) – can help attenuate the link between the marginal price setter and the cost of energy for end users, but such solutions are underdeveloped in Europe, in turn limiting the benefits from accelerating the roll-out of renewables. In the absence of action, this decoupling problem will remain acute at least for the remainder of this decade. Even if renewable installation targets are met, it is not forecast to significantly reduce the share of hours during which fossil fuels set energy prices by 2030.

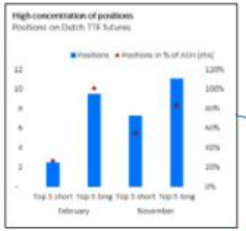
Text embedding model

Chart caption

Note: Absolute value of net positions in EUR billion for the top five long and short non-financial corporate counterparties and positions in % of average daily trading volume, in % rhs. The high concentration of positions indicates that if several firms with similar directional positions were to reduce their exposures, they could amplify market moves.

Sources: EMIR, ESMA

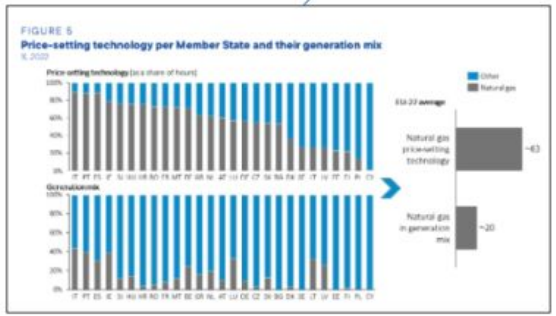
Chart



Good visual embedding model for charts(??)

Combining everything into one meaningful representation

Figure



Layout

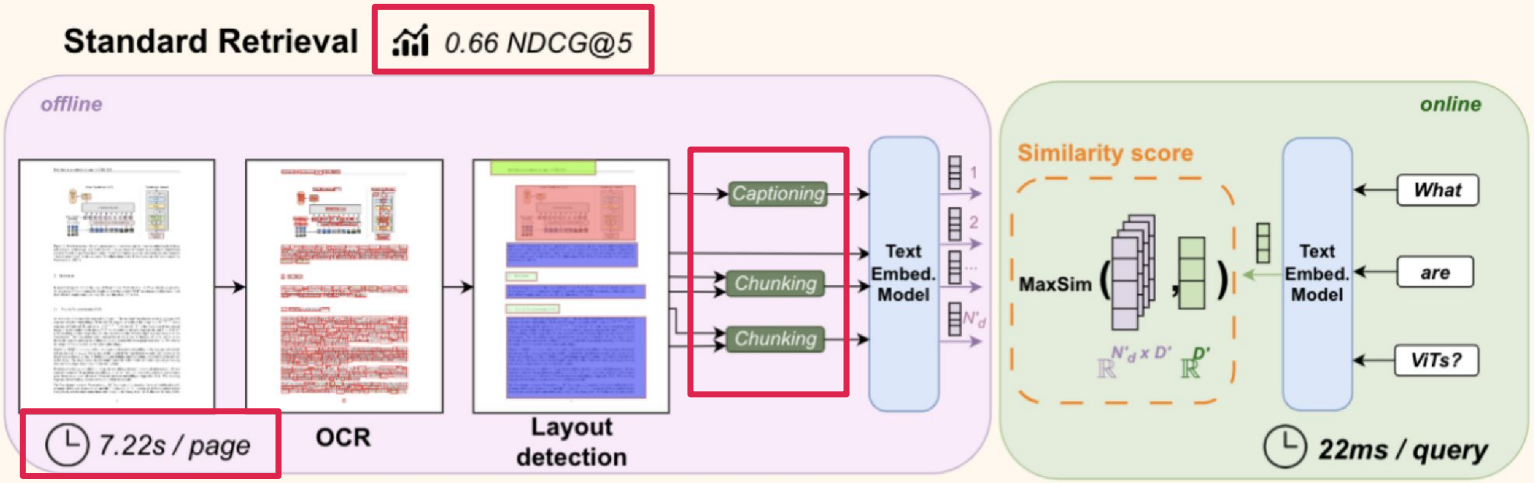
Chart

Chart caption

Text

Figure

Motivation to Move Beyond OCR-Based Methods*



Adapted from "ColPali: Efficient Document Retrieval with Vision Language Models," by Faysse et al., 2024.

*According to the authors of the Visual Language Model-based PDF retrieval (June, 2024).

My (Side Note on) Motivation

OCR is evolving — on March 6, Mistral OCR launched:

- ~2,000 pages per minute (for ~2$)
- Strong benchmarks (claimed to be better than GPT-4o)

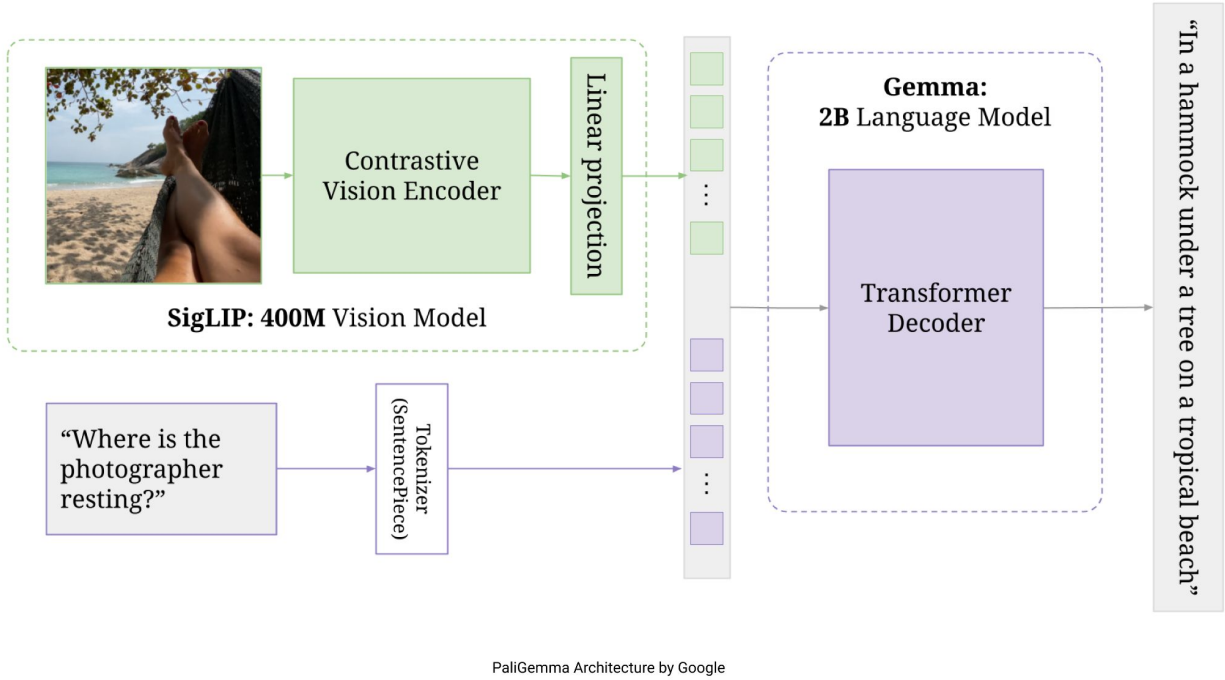
*Imo, the main reason to use VLMs for PDF retrieval is to **avoid the challenge** of splitting, embedding, and combining data from PDF pages, so this retrieval works.*

Visual Language Models in PDF retrieval

Vision Language Models (VLMs)

VLMs process both **images** and **text** as inputs, encoding them into a **shared embedding space** (CLIP, Llava, Qwen2, Blip, PaliGemma3, etc.)

The authors of the Visual Language Model-based PDF retrieval method built this method on **PaliGemma-3B**:

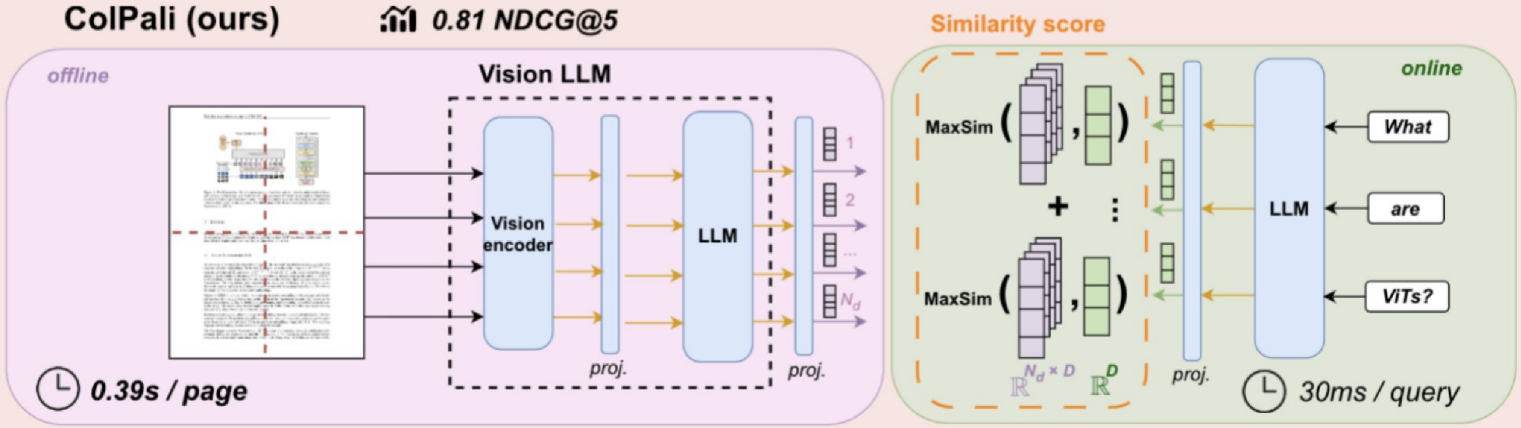


PaliGemma Architecture by Google

Contextualized Late Interaction over PaliGemma

PDF pages (images) are split into small image patches.
Queries (texts) are split into tokens.
Let's call patches & tokens ***'pieces***.'

1. Use cross-attended and aligned piece **embeddings** from **PaliGemma-3B** (remove the generative layer(s)).
2. **Project** them **down** to make them lighter.
3. Use a **late interaction mechanism**, aka train to estimate similarity **between** pages and queries **pieces**.

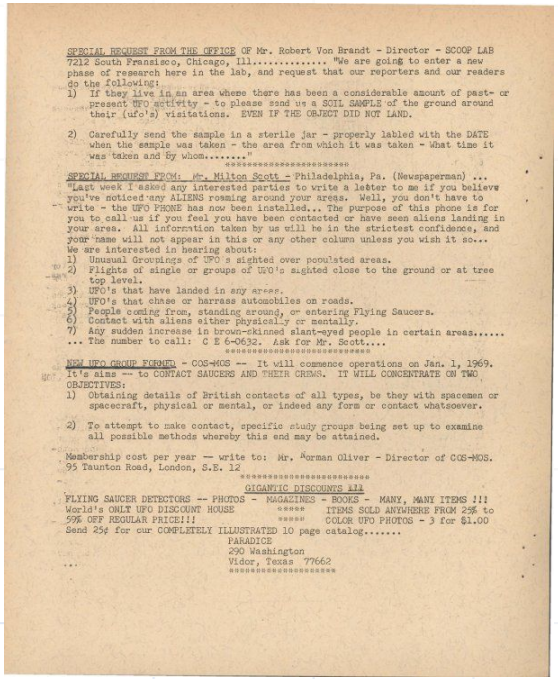


Adapted from "ColPali: Efficient Document Retrieval with Vision Language Models," by Faysse et al., 2024.

ColPali for Retrieval: Step by Step

Offline Phase: Store PDFs

Each document page is divided into a 32×32 grid, creating 1024 patches (+ some postfix, aka “describe the image”)



SPECIAL REQUEST FROM THE OFFICE OF Mr. Robert Von Brandt - Director - SCOOP LAB
7212 South Fransisco, Chicago, Ill..... "We are going to enter a new
phase of research here in the lab, and request that our reporters and our readers
do the following:
1) If they live in an area where there has been a considerable amount of past- or
present UFO activity - to please send us a SOIL SAMPLE of the ground around
their (ufo's) visitations. EVEN IF THE OBJECT DID NOT LAND.

2) Carefully send the sample in a sterile jar - properly labled with the DATE
when the sample was taken - the area from which it was taken - what time it
was taken and by whom....."
****

SPECIAL REQUEST FROM: Mr. Milton Scott - Philadelphia, Pa. (Newspaperman) ...
"Last week I asked any interested parties to write a letter to me if you believe
you've noticed any ALIENS roaming around your areas. Well, you don't have to
write - the UFO PHONE has now been installed... The purpose of this phone is for
you to call us if you feel you have been contacted or have seen aliens landing in
your area. All information taken by us will be in the strictest confidence, and
your name will not appear in this or any other column unless you wish it so...
We are interested in hearing about:
1) Unusual Groupings of UFO's sighted over populated areas.
2) Flights of single or groups of UFO's sighted close to the ground or at tree
top level.
3) UFO's that have landed in any areas.
4) UFO's that chase or harrass automobiles on roads.
5) People coming from, standing around, or entering Flying Saucers.
6) Contact with aliens either physical or mentally.
7) Any sudden increase in brown-skinned slant-eyed people in certain areas.....
... The number to call: C E 6-0632. Ask for Mr. Scott....
****

NEW UFO GROUP FORMED - COS-MOS -- It will commence operations on Jan. 1, 1969.
It's aims -- to CONTACT SAUCERS AND THEIR CREWS. IT WILL CONCENTRATE ON TWO
OBJECTIVES:
1) Obtaining details of British contacts of all types, be they with spacemen or
spacecraft, physical or mental, or indeed any form or contact whatsoever.

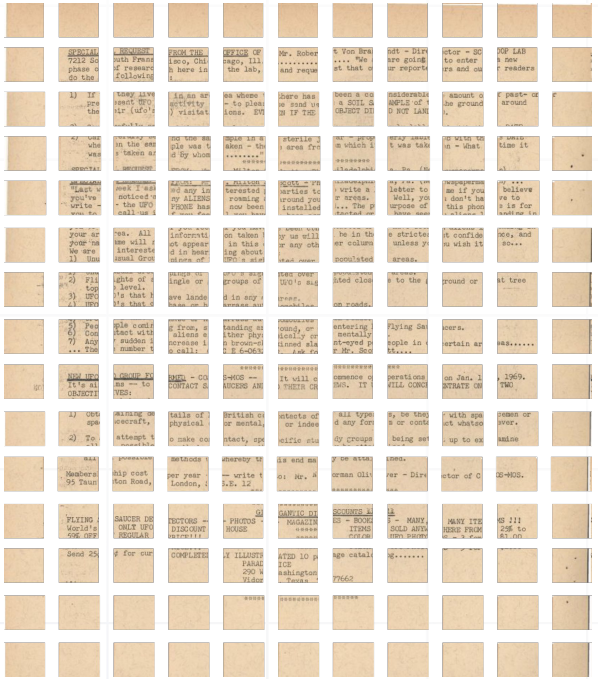
2) To attempt to make contact, specific study groups being set up to examine
all possible methods whereby this end may be attained.

Membership cost per year -- write to: Mr. Norman Oliver - Director of COS-MOS.
95 Taunton Road, London, S.E. 12
****

GIGANTIC DISCOUNTS !!!

FLYING SAUCER DETECTORS -- PHOTOS - MAGAZINES - BOOKS - MANY, MANY ITEMS !!!
World's ONLY UFO DISCOUNT HOUSE ***** ITEMS SOLD ANYWHERE FROM 25% to
59% OFF REGULAR PRICE!!! ***** COLOR UFO PHOTOS - 3 for $1.00
Send 25¢ for our COMPLETELY ILLUSTRATED 10 page catalog.....
PARADICE
290 Washington
Vidor, Texas 77662
****

PDF page (image)



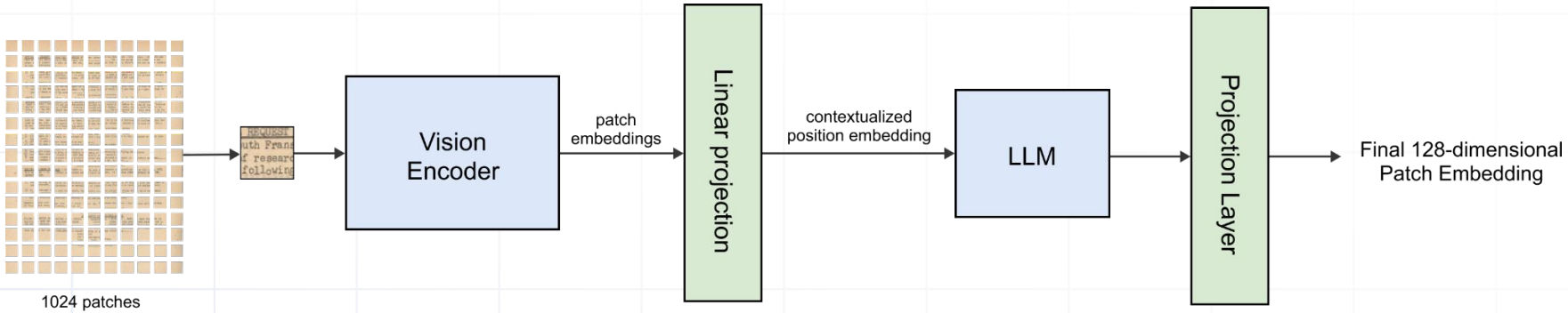
Patched image

Offline Phase: Store PDFs

Generate multivector embeddings for the entire page (per patch) & store in a vector database:

- Captures local visual features from patches.
- Understands the relative positions of these features on the page.
- Aligns the visual and textual embeddings into a shared embedding space for retrieval.

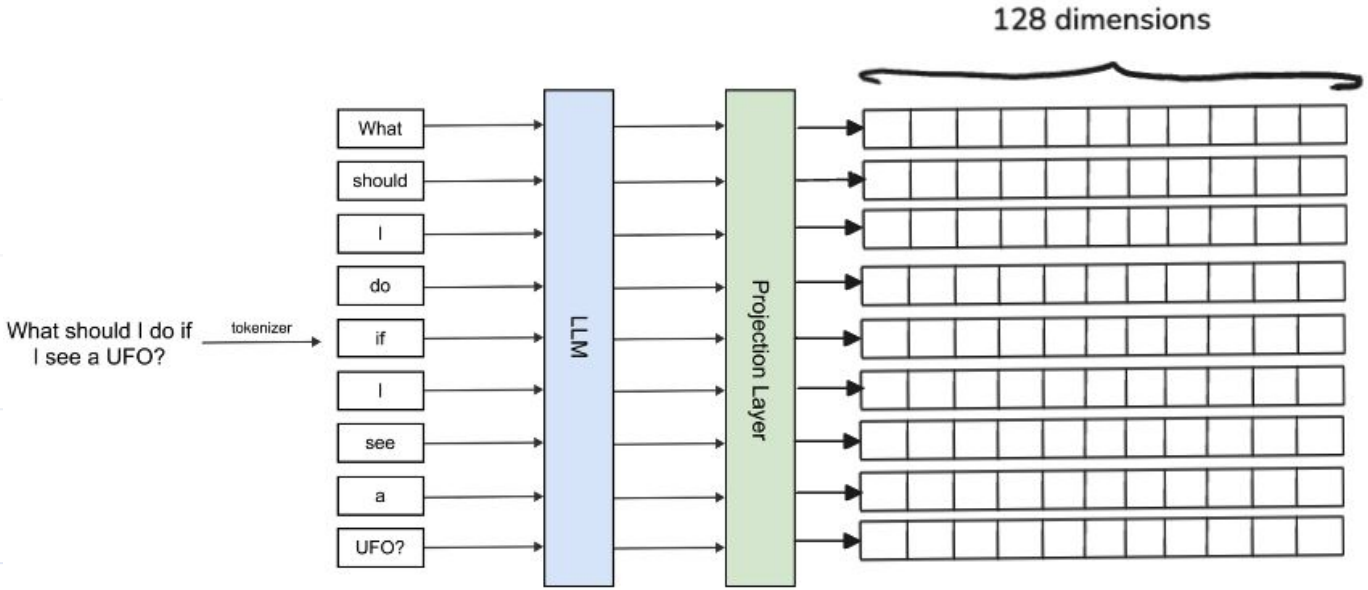
1 patch = 128-dimensional vector



Online phase: Query processing

To start the retrieval process, we need to process our query, tokenize it and embed each token into the same space as PDF document patches.

1 token = 128-dimensional vector



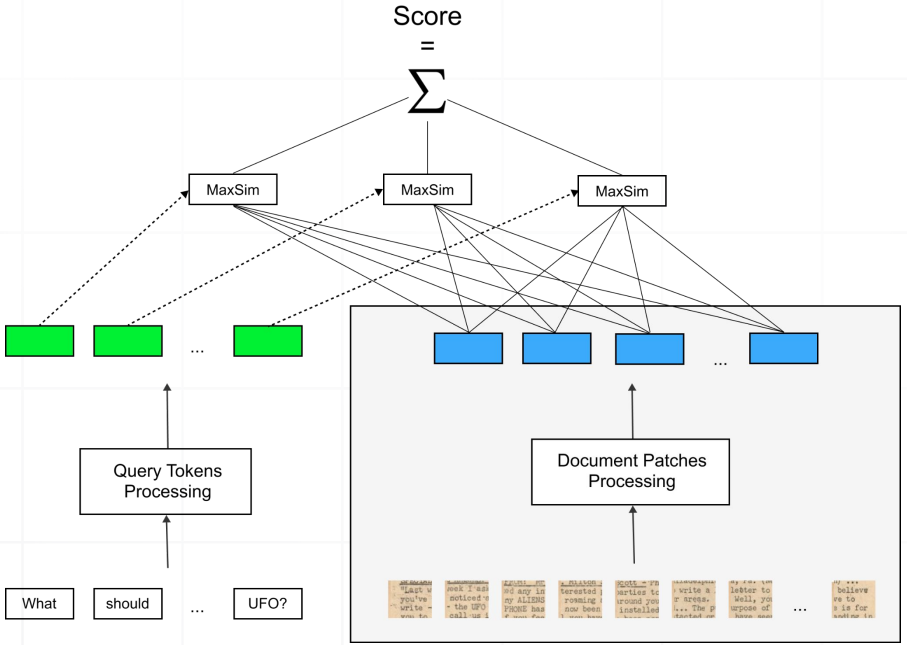
Online phase: Contextualized Late Interaction (CoI)

At retrieval time, **MaxSim pooling** is used to compute the similarity scores between all **query** and **document pieces**, pair by pair.

The **similarity** is defined as **cosine** between 128-dimensional vectors-pieces.

1 query token "finds" the maximally similar page piece, and this similarity score is saved for this token.

The final relevance score is computed by **summing the MaxSim scores for all query tokens.**



Evaluating ColPali, The ViDoRe Benchmark

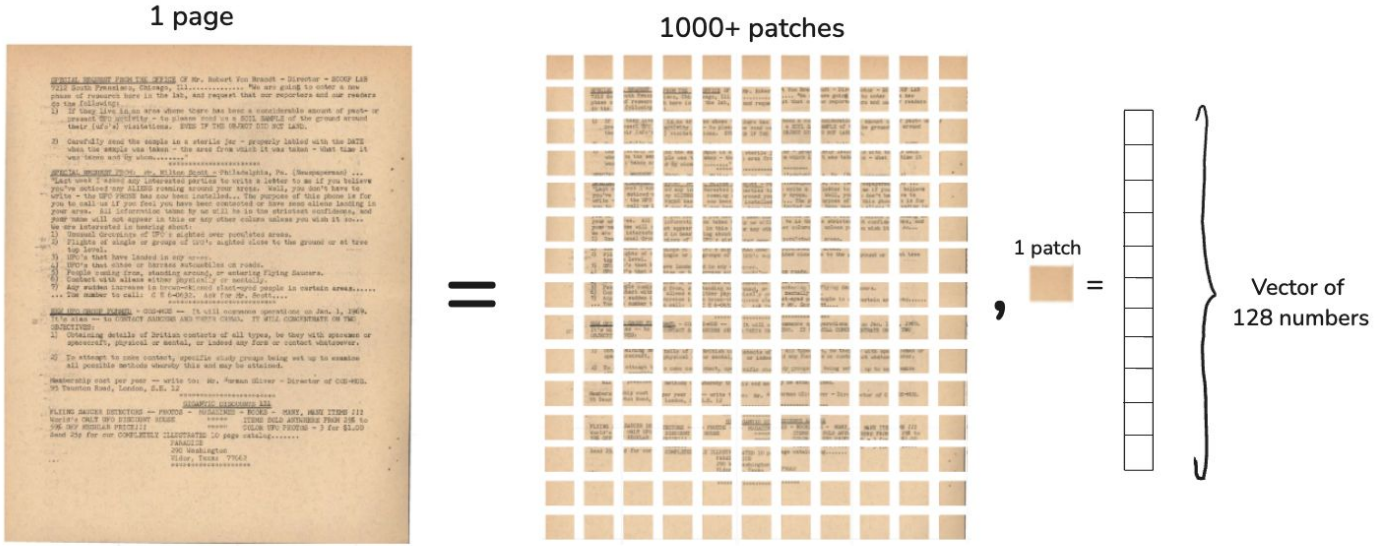
	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health.	Avg.
Unstructured Text only											
- BM25	-	34.1	-	-	44.0	59.6	90.4	78.3	78.8	82.6	-
- BGE-M3	-	28.4↓5.7	-	-	36.1↓7.9	68.5↑8.9	88.4↓2.0	76.8↓1.5	77.7↓1.1	84.6↑2.0	-
Unstructured + OCR											
- BM25	31.6	36.8	62.9	46.5	62.7	64.3	92.8	85.9	83.9	87.2	65.5
- BGE-M3	31.4↓0.2	25.7↓11.1	60.1↓2.8	70.8↑24.3	50.5↓12.2	**73.2**↑8.9	90.2↓2.6	83.6↓2.3	84.9↑1.0	91.1↑3.9	66.1↑0.6
Unstructured + Captioning											
- BM25	40.1	38.4	70.0	35.4	61.5	60.9	88.0	84.7	82.7	89.2	65.1
- BGE-M3	35.7↓4.4	32.9↓5.4	71.9↑1.9	69.1↑33.7	43.8↓17.7	73.1↑12.2	88.8↑0.8	83.3↓1.4	80.4↓2.3	91.3↑2.1	67.0↑1.9
Contrastive VLMs											
Jina-CLIP	25.4	11.9	35.5	20.2	3.3	3.8	15.2	19.7	21.4	20.8	17.7
Nomic-vision	17.1	10.7	30.1	16.3	2.7	1.1	12.9	10.9	11.4	15.7	12.9
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
Ours											
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
BiSigLIP (+fine-tuning)	58.5↑15.3	32.9↑2.6	70.5↑6.4	62.7↑4.6	30.5↑4.3	26.5↑7.8	74.3↑11.8	73.7↑8.0	74.2↑8.1	82.3↑3.2	58.6↑7.2
BiPali (+LLM)	56.5↓-2.0	30.0↓-2.9	67.4↓-3.1	76.9↑14.2	33.4↑2.9	43.7↑17.2	71.2↓-3.1	61.9↓-11.7	73.8↓-0.4	73.6↓-8.8	58.8↑0.2
ColPali (+Late Inter.)	**79.1**↑22.6	**54.4**↑24.5	**81.8**↑14.4	**83.9**↑7.0	**65.8**↑32.4	**73.2**↑29.5	**96.2**↑25.0	**91.0**↑29.1	**92.7**↑18.9	**94.4**↑20.8	**81.3**↑22.5

Table 2: **Comprehensive evaluation of baseline models and our proposed method on *ViDoRe*.** Results are presented using NDCG@5 metrics, and illustrate the impact of different components. Text-only metrics are not computed for benchmarks with only visual elements.

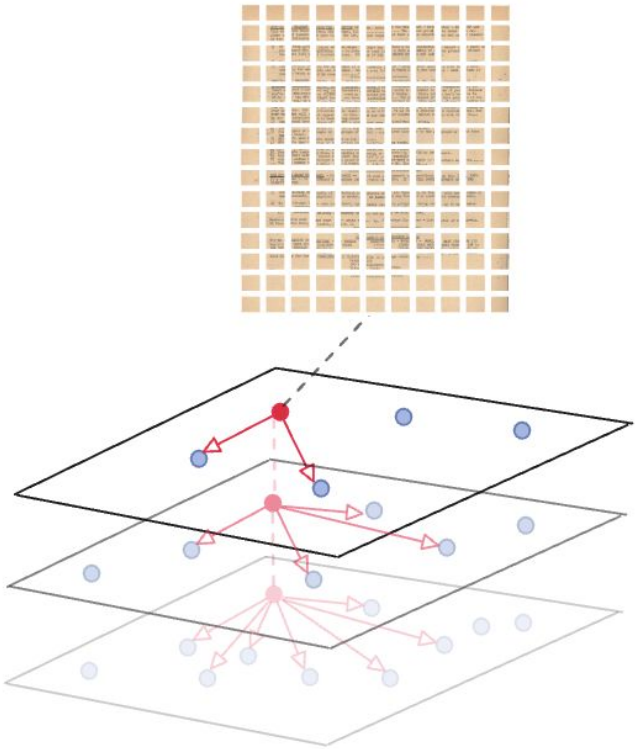
And Then People Tried to Use It

Math Behind Scaling

Let’s dump **1000+** 128 dimensional vectors **per PDF page** in a vector database which supports multivectors and see what happens.



Math Behind Scaling



Inserting a point into a popular vector index (HNSW)*

To simplify:
Imagine you're comparing one point with 100 others during index construction.

- 1 point = 1000 vectors
- Another point = 1000 vectors
- 100 × 1000 × 1000 = **100 million vector comparisons**

For just **ONE page**.

* A vector index organizes vector representations in a database, enabling fast searches at scale (billions of vectors).

So, We Tried to Fix it

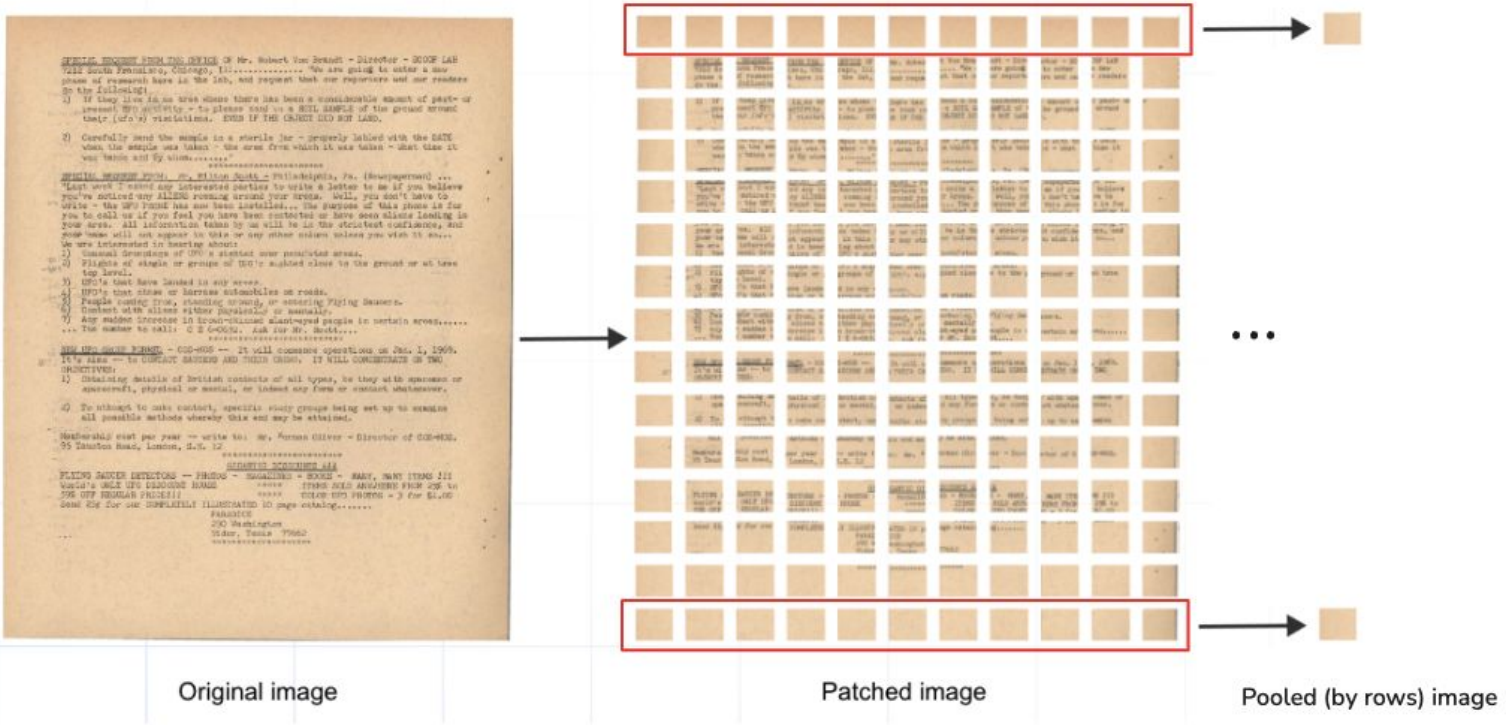
Idea

100 × 1000 × 1000 = 100 million vector comparisons, just to organize vectors within a database in a structure that allows fast search.

We need to **reduce the number of comparisons without losing search quality.**

Also since these comparisons affect retrieval speed — since, as you remember, a query is also a bunch of vectors that need to be compared to pages.

Idea: Pooling Image Patches

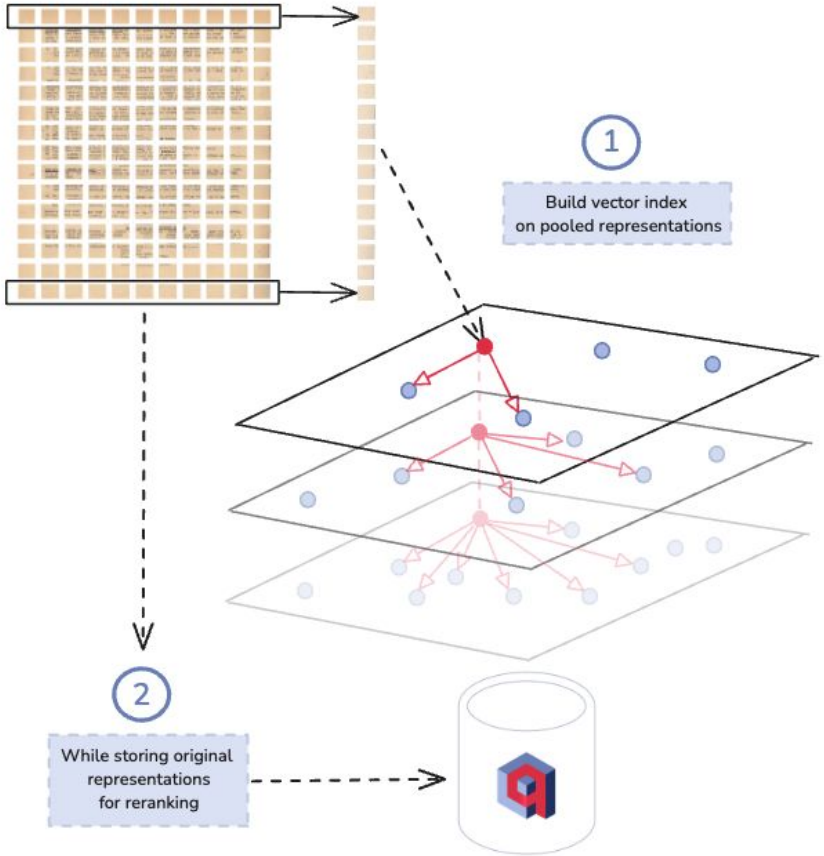


Original image

Patched image

Pooled (by rows) image

Experiment Idea

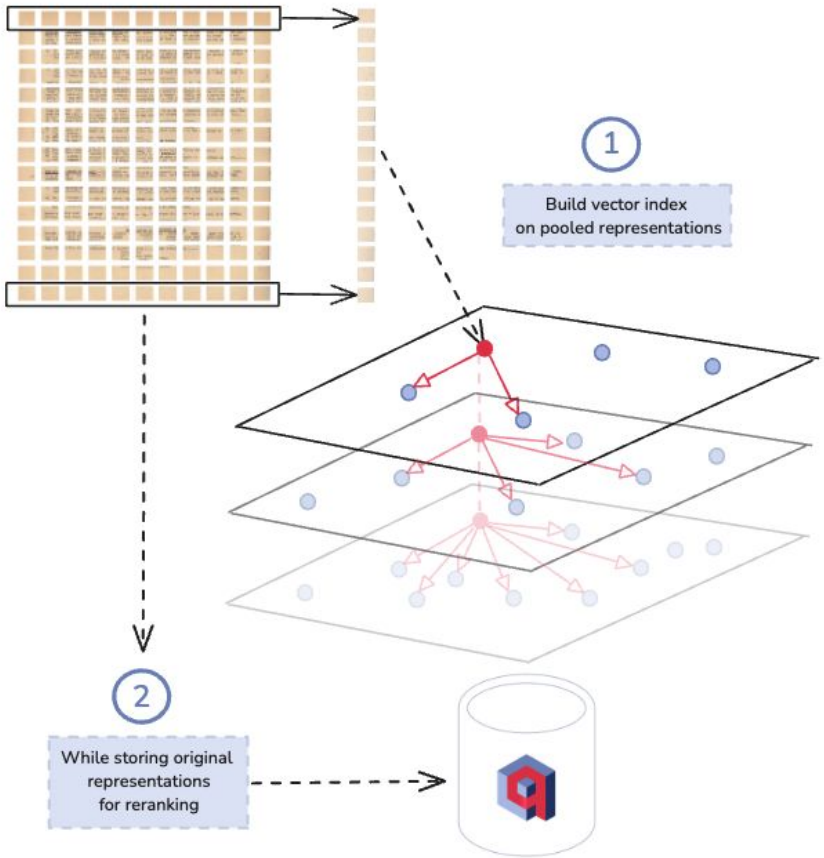


1. Test pooling of patches, pool page row (or column) vectors into one;
2. Build vector index on them;
3. Save the original ColPali representations per page for reranking.

Compare (pace & quality):

1. **Original** ColPali-based retrieval;
2. **Retrieving** X results based on pooled vectors + **reranking** them with the original representations.

Experiment Setup



1. Test **mean/max** pooling of patches, pool page **row** vectors into one;
2. Build vector index on them;
3. Save the original ColPali representations per page for reranking.

Compare (pace & quality)

1. **Original** ColPali-based Retrieval;
2. **Retrieving 200** results based on pooled vectors + **reranking** them with the original representations.

On:

- **21k+** pages;
- **1000** random queries;
- **NDCG@20**.

Experiment Results

- **13x faster** than the original retrieval
- **NDCG@20 = 0.95** compared to retrieval with original vectors.
- **Mean pooling works**, while max pooling didn't preserve retrieval quality.

Also, I experimented with cutting query padding tokens & binary quantization — you can check it out in the repo:)

GitHub repo with experiment

QRs for the QR god

Official PDF Retrieval Scaling Guide

Col-VLMs

After **ColPali**, **ColQwen2** (based on Qwen2) was released:

- Dynamic page resizing (patching PDF pages dynamically)
- Better results on ViDoRe (+6% NDCG@5)
- Fewer vectors (~750 instead of 1000+ per page)

Since its embedding representations are slightly different, community started asking how to use it in production.

So, I wrote an **official code guide on Col-VLMs for PDF retrieval**.

Official Guide on Col-VLMs for PDF Retrieval

1. Construct an HNSW index using only mean-pooled vectors. Use them for the **first-stage retrieval**.

2. Use the original multivectors from ColPali or ColQwen2 to **rerank** the results retrieved in the first stage.

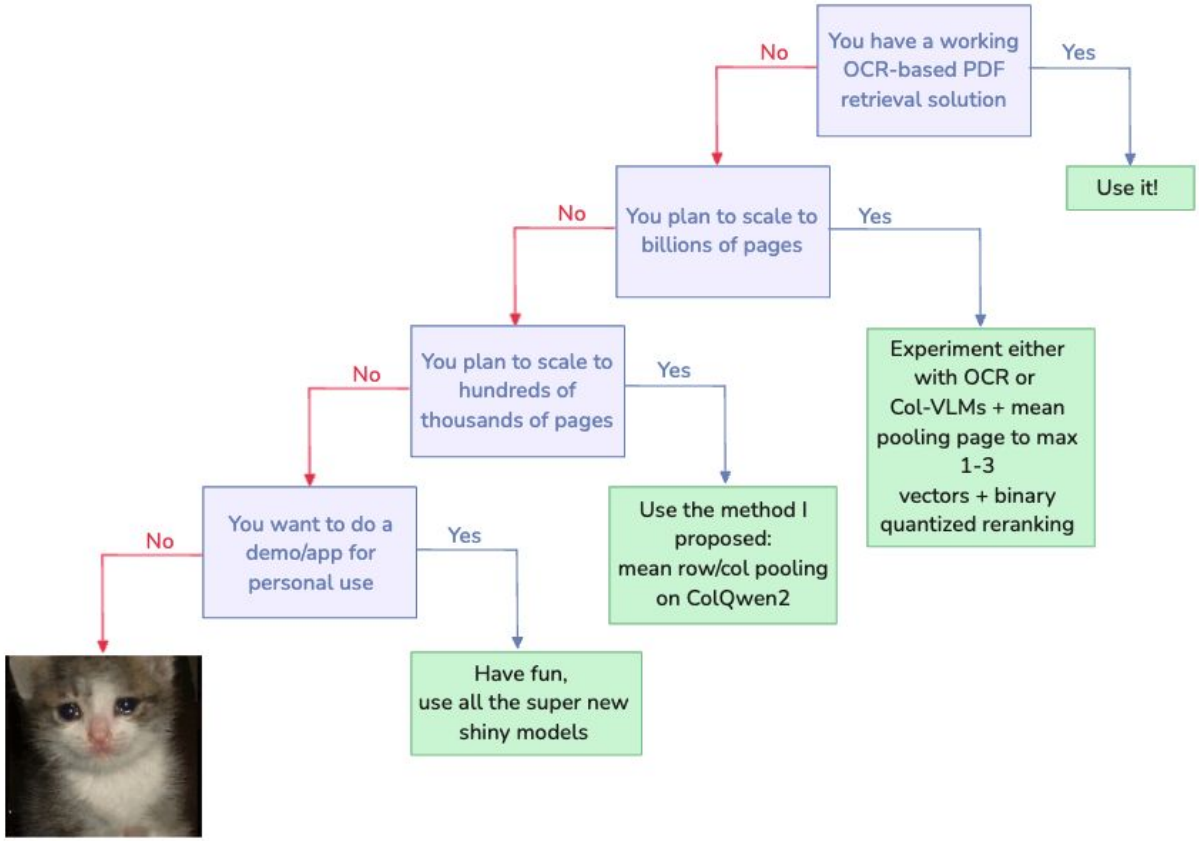
- Experiment with numbers for reranking;
- Use pooled by rows/by columns or both;
- Experiment with quantization and vector precision

Guide for ColPali/ColQwen2

QRs for the QR god

Takeaways

Main (Mine) Takeaways

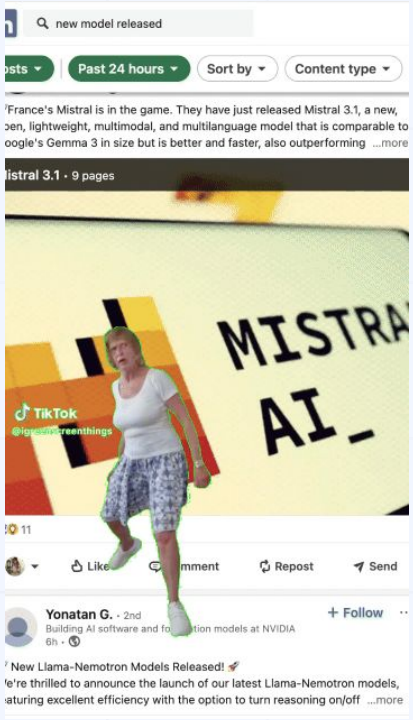


What about <insert model name>?

- **SmolDocLink** (March 14, 2025) – interesting PDF-like document representations;
- **Gemma 3** (March 12, 2025) – big smart multilingual VLM;
- **Mistral OCR** (March 6, 2025) – latest OCR;

...

At the end of the day, **use whatever works for you** — just **don't forget** about **cost, time and quality** at **scale**.



Time for Q&A!

For everything else, there's my LinkedIn;)