# What You See Is What You Search: Vision Language Models for PDF Retrieval

Jo Kristian Bergum, Vespa.ai

30.09.2024

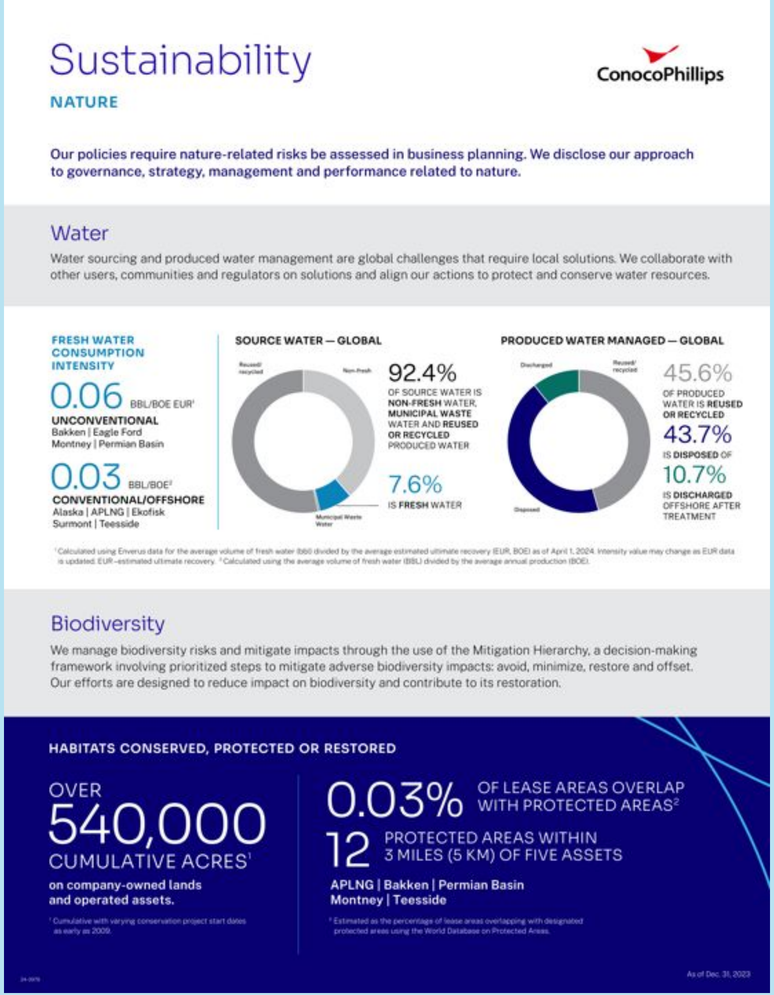
Vespa.ai

# Agenda

1. **Introduction (where we are now)**
2. Indexing & Searching PDFs
3. Vision Language Models (VLMs)
4. ColPali - Using VLM for retrieval
5. Scaling ColPali to large collections
6. RAG with VLMs
7. QA & Resources

This deck:

# Indexing & Searching in PDFs


ConocoPhillips
Reused/recycled
Non-Fresh
Municipal Waste Water
Discharged
Reused/recycled
Disposed

## Sustainability

**NATURE**

Our policies require nature-related risks be assessed in business planning. We disclose our approach to governance, strategy, management and performance related to nature.

### Water

Water sourcing and produced water management are global challenges that require local solutions. We collaborate with other users, communities and regulators on solutions and align our actions to protect and conserve water resources.

**FRESH WATER CONSUMPTION INTENSITY**

0.06 BBL/BOE EUR[1]
**UNCONVENTIONAL**
Bakken | Eagle Ford
Montney | Permian Basin

0.03 BBL/BOE[2]
**CONVENTIONAL/OFFSHORE**
Alaska | APLNG | Ekofisk
Surmont | Teesside

**SOURCE WATER — GLOBAL**

92.4%
OF SOURCE WATER IS **NON-FRESH** WATER, **MUNICIPAL WASTE** WATER AND **REUSED OR RECYCLED** PRODUCED WATER

7.6%
IS **FRESH** WATER

**PRODUCED WATER MANAGED — GLOBAL**

45.6%
OF PRODUCED WATER IS **REUSED OR RECYCLED**

43.7%
IS **DISPOSED** OF

10.7%
IS **DISCHARGED** OFFSHORE AFTER TREATMENT

[1] Calculated using Enverus data for the average volume of fresh water (bbl) divided by the average estimated ultimate recovery (EUR, BOE) as of April 1, 2024. Intensity value may change as EUR data is updated. EUR–estimated ultimate recovery. [2] Calculated using the average volume of fresh water (BBL) divided by the average annual production (BOE).

### Biodiversity

We manage biodiversity risks and mitigate impacts through the use of the Mitigation Hierarchy, a decision-making framework involving prioritized steps to mitigate adverse biodiversity impacts: avoid, minimize, restore and offset. Our efforts are designed to reduce impact on biodiversity and contribute to its restoration.

**HABITATS CONSERVED, PROTECTED OR RESTORED**

OVER
540,000
CUMULATIVE ACRES[1]
**on company-owned lands and operated assets.**

[1] Cumulative with varying conservation project start dates as early as 2009.

0.03% OF LEASE AREAS OVERLAP WITH PROTECTED AREAS[2]

12 PROTECTED AREAS WITHIN 3 MILES (5 KM) OF FIVE ASSETS

**APLNG | Bakken | Permian Basin**
**Montney | Teesside**

[2] Estimated as the percentage of lease areas overlapping with designated protected areas using the World Database on Protected Areas.

As of Dec. 31, 2023

Complex format with visual infographics, tables, charts and images.

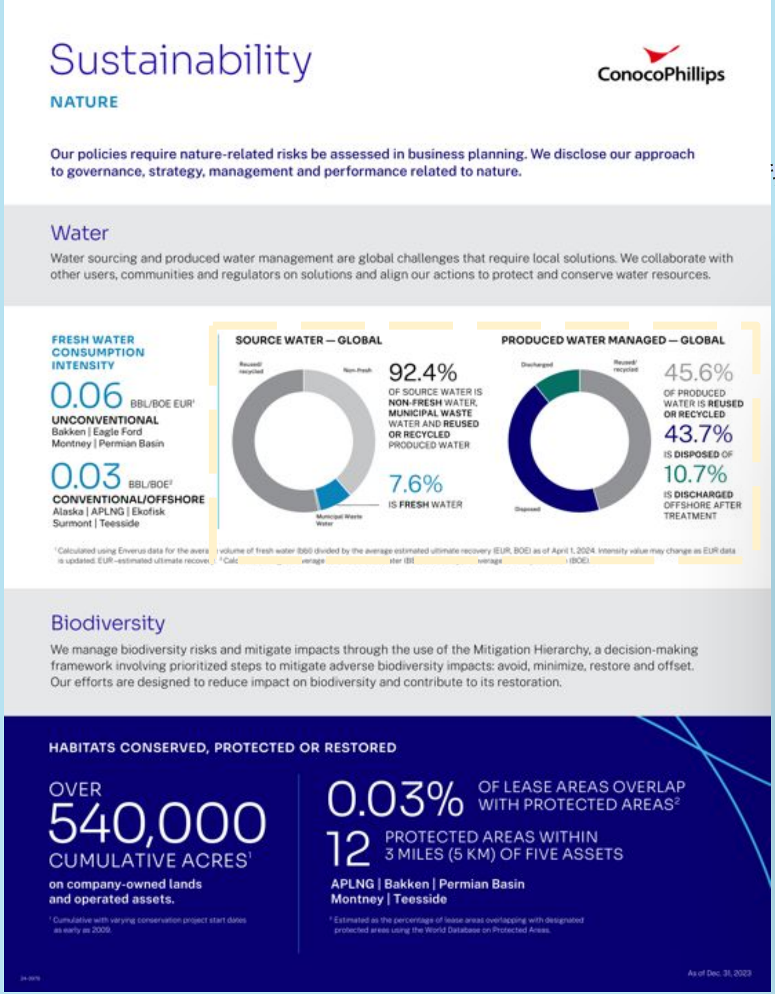
Searching PDFs was Popularized by RAG (Retrieval-Augmented Generation) - Chat with your PDFs

Extraction pipeline to map to text:

- Chunking?
- Text Embedding?

# Text extraction is *messy*

- Embedded images (e.g. infographics)
- Order
- Layout & Structure
- Tables
- Font size



## Sustainability

**NATURE**

ConocoPhillips

**Our policies require nature-related risks be assessed in business planning. We disclose our approach to governance, strategy, management and performance related to nature.**

### Water

Water sourcing and produced water management are global challenges that require local solutions. We collaborate with other users, communities and regulators on solutions and align our actions to protect and conserve water resources.

**FRESH WATER CONSUMPTION INTENSITY**

0.06 BBL/BOE EUR[1]
**UNCONVENTIONAL**
Bakken | Eagle Ford
Montney | Permian Basin

0.03 BBL/BOE[2]
**CONVENTIONAL/OFFSHORE**
Alaska | APLNG | Ekofisk
Surmont | Teesside

**SOURCE WATER — GLOBAL**

92.4% OF SOURCE WATER IS **NON-FRESH** WATER, **MUNICIPAL WASTE** WATER AND **REUSED OR RECYCLED** PRODUCED WATER

7.6% IS **FRESH** WATER

**PRODUCED WATER MANAGED — GLOBAL**

45.6% OF PRODUCED WATER IS **REUSED OR RECYCLED**

43.7% IS **DISPOSED** OF

10.7% IS **DISCHARGED** OFFSHORE AFTER TREATMENT

### Biodiversity

We manage biodiversity risks and mitigate impacts through the use of the Mitigation Hierarchy, a decision-making framework involving prioritized steps to mitigate adverse biodiversity impacts: avoid, minimize, restore and offset. Our efforts are designed to reduce impact on biodiversity and contribute to its restoration.

**HABITATS CONSERVED, PROTECTED OR RESTORED**

OVER
540,000
CUMULATIVE ACRES[1]
**on company-owned lands and operated assets.**

[1] Cumulative with varying conservation project start dates as early as 2009.

0.03% OF LEASE AREAS OVERLAP WITH PROTECTED AREAS[2]

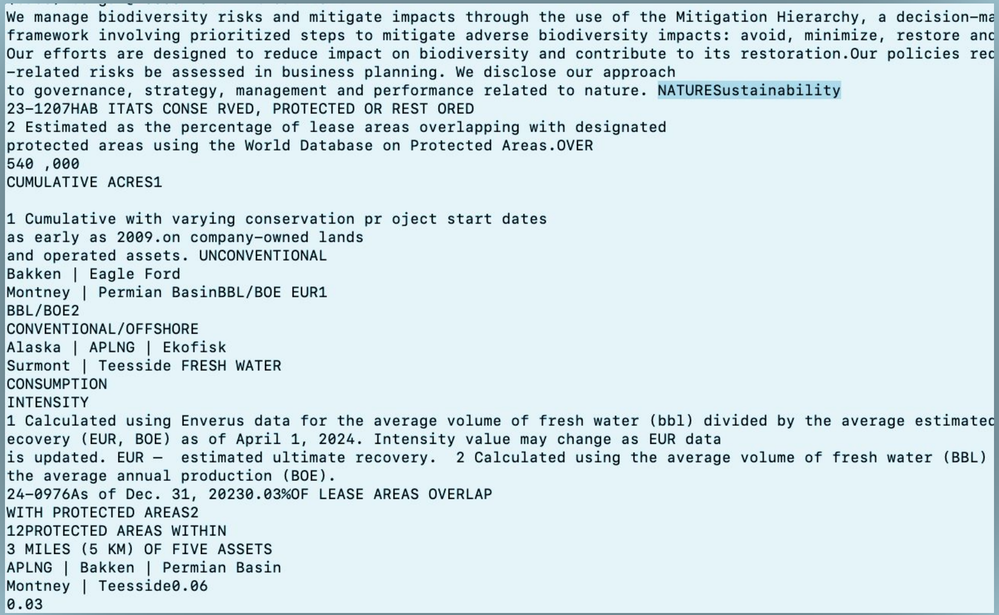
12 PROTECTED AREAS WITHIN 3 MILES (5 KM) OF FIVE ASSETS

**APLNG | Bakken | Permian Basin**
**Montney | Teesside**

[2] Estimated as the percentage of lease areas overlapping with designated protected areas using the World Database on Protected Areas.

As of Dec. 31, 2023

Extract



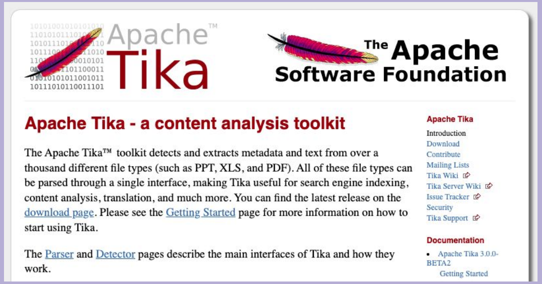
```
We manage biodiversity risks and mitigate impacts through the use of the Mitigation Hierarchy, a decision-ma
framework involving prioritized steps to mitigate adverse biodiversity impacts: avoid, minimize, restore and
Our efforts are designed to reduce impact on biodiversity and contribute to its restoration.Our policies rec
-related risks be assessed in business planning. We disclose our approach
to governance, strategy, management and performance related to nature. NATURESustainability
23-1207HAB ITATS CONSE RVED, PROTECTED OR REST ORED
2 Estimated as the percentage of lease areas overlapping with designated
protected areas using the World Database on Protected Areas.OVER
540 ,000
CUMULATIVE ACRES1

1 Cumulative with varying conservation pr oject start dates
as early as 2009.on company-owned lands
and operated assets. UNCONVENTIONAL
Bakken | Eagle Ford
Montney | Permian BasinBBL/BOE EUR1
BBL/BOE2
CONVENTIONAL/OFFSHORE
Alaska | APLNG | Ekofisk
Surmont | Teesside FRESH WATER
CONSUMPTION
INTENSITY
1 Calculated using Enverus data for the average volume of fresh water (bbl) divided by the average estimatec
ecovery (EUR, BOE) as of April 1, 2024. Intensity value may change as EUR data
is updated. EUR –  estimated ultimate recovery.  2 Calculated using the average volume of fresh water (BBL)
the average annual production (BOE).
24-0976As of Dec. 31, 20230.03%OF LEASE AREAS OVERLAP
WITH PROTECTED AREAS2
12PROTECTED AREAS WITHIN
3 MILES (5 KM) OF FIVE ASSETS
APLNG | Bakken | Permian Basin
Montney | Teesside0.06
0.03
```

# *Why extract to text in the first place?*

*Convert to something that makes it searchable*

Convert to text modality:

=> Text search
- Keyword search (BM25)
- Embedding retrieval with text embedding models (e.g. Jina)
- Splade, Bert, ColBERT, OpenAi bla bla bla, all text-only



Apache Tika - a content analysis toolkit

The Apache Tika™ toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF). All of these file types can be parsed through a single interface, making Tika useful for search engine indexing, content analysis, translation, and much more. You can find the latest release on the download page. Please see the Getting Started page for more information on how to start using Tika.

The Parser and Detector pages describe the main interfaces of Tika and how they work.

UNSTRUCTURED

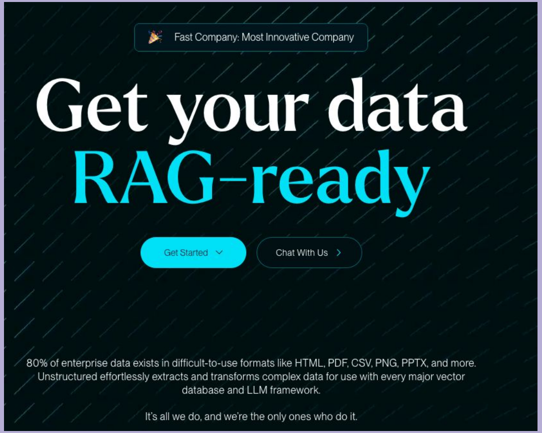
LlamaCloud

Amazon Textract

Azure AI Document Intelligence

Accelerate information extraction from documents.

Tesseract



Get your data RAG-ready

80% of enterprise data exists in difficult-to-use formats like HTML, PDF, CSV, PNG, PPTX, and more. Unstructured effortlessly extracts and transforms complex data for use with every major vector database and LLM framework.

It's all we do, and we're the only ones who do it.

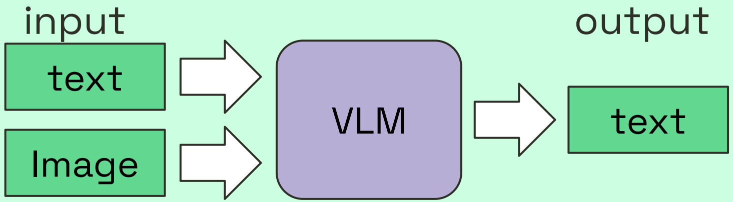
# Agenda

1. ~~Introduction (where we are now)~~
2. ~~Indexing & Searching PDFs~~
3. **Vision Language Models (VLMs)**
4. ColPali - Using VLM for retrieval
5. Scaling ColPali to large collections
6. RAG with VLMs
7. QA & Resources

# What is a Vision Language Model (VLM) ?

*Vision language models are broadly defined as multimodal models that can learn from images and text.*

*They are a type of generative models that take image and text inputs, and generate text output*



input

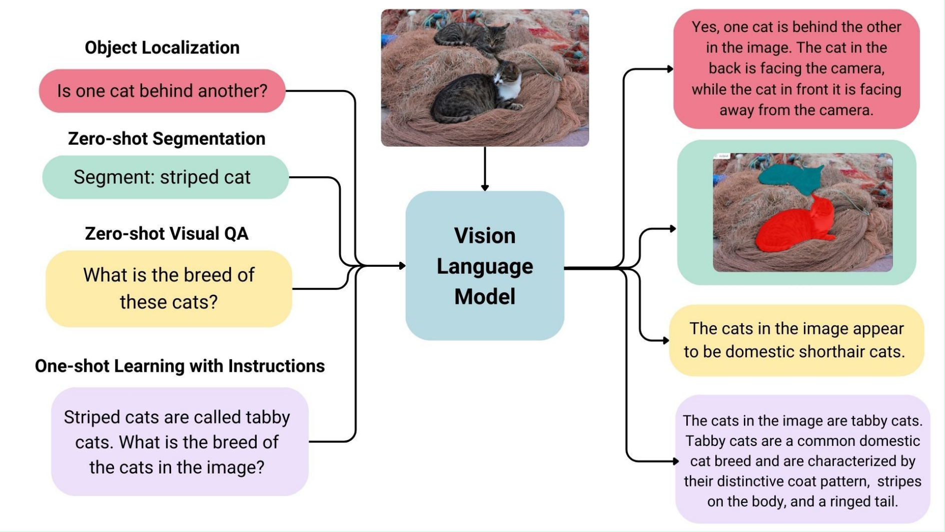
text

Image

VLM

output

text



**Object Localization**

Is one cat behind another?

**Zero-shot Segmentation**

Segment: striped cat

**Zero-shot Visual QA**

What is the breed of these cats?

**One-shot Learning with Instructions**

Striped cats are called tabby cats. What is the breed of the cats in the image?

**Vision Language Model**

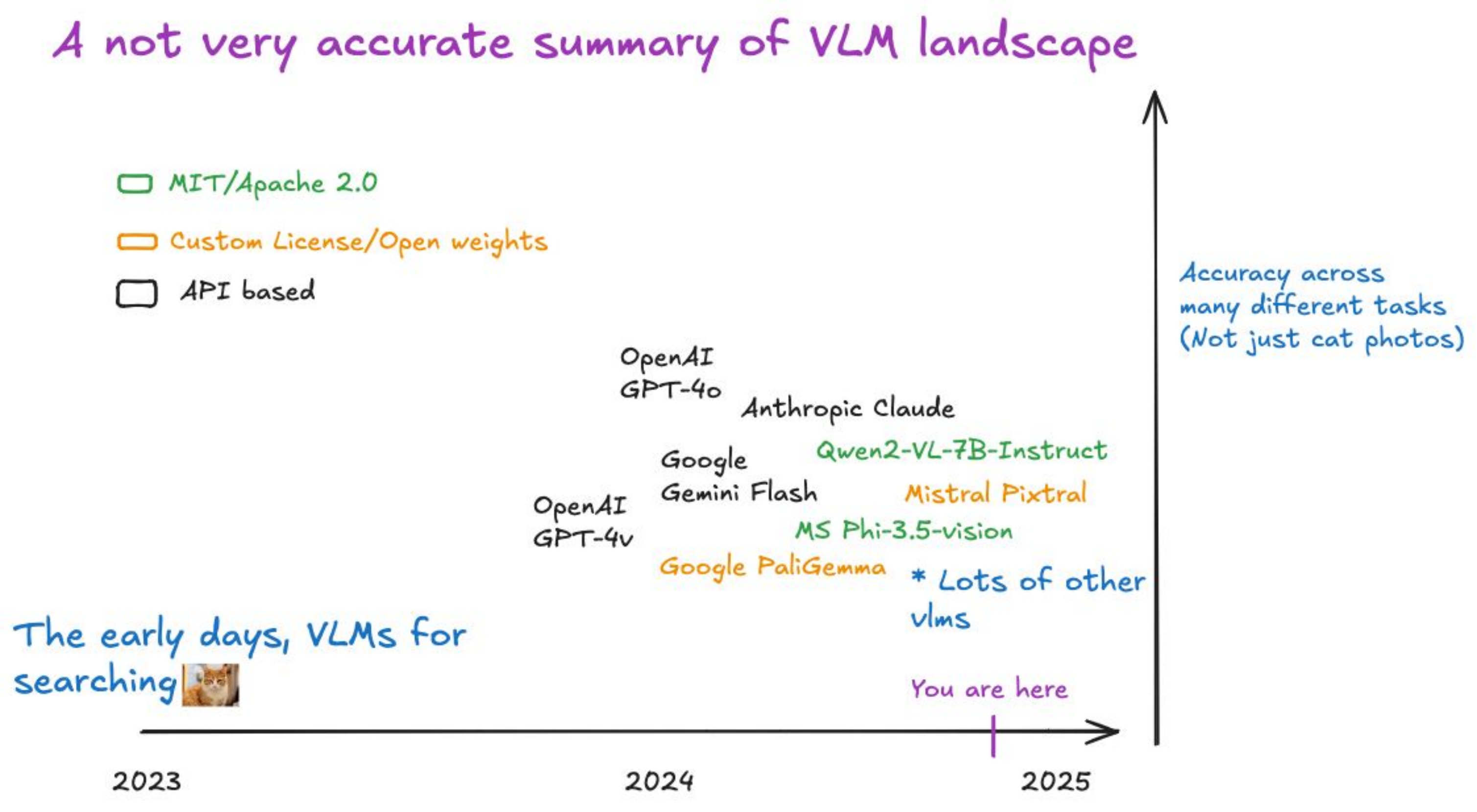
Yes, one cat is behind the other in the image. The cat in the back is facing the camera, while the cat in front it is facing away from the camera.

The cats in the image appear to be domestic shorthair cats.

The cats in the image are tabby cats. Tabby cats are a common domestic cat breed and are characterized by their distinctive coat pattern, stripes on the body, and a ringed tail.

*From [https://huggingface.co/blog/vlms](https://huggingface.co/blog/vlms)*

# The year of multimodal AI (ok, VLMs)



New VLMs last week

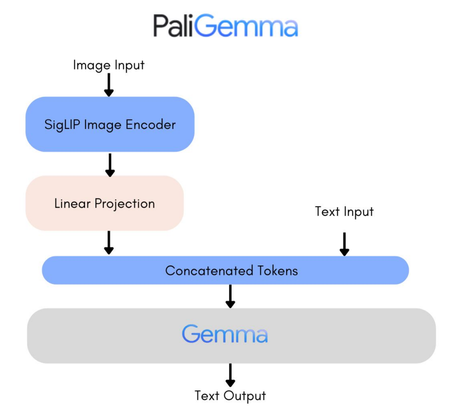
Molmo

Llama 3.2

# PaliGemma (a VLM)

PaliGemma – Google's *"Cutting-Edge"* Open Weight VLM

-OCR capabilities
- Question answering + more



PaliGemma

Image Input

SigLIP Image Encoder

Linear Projection

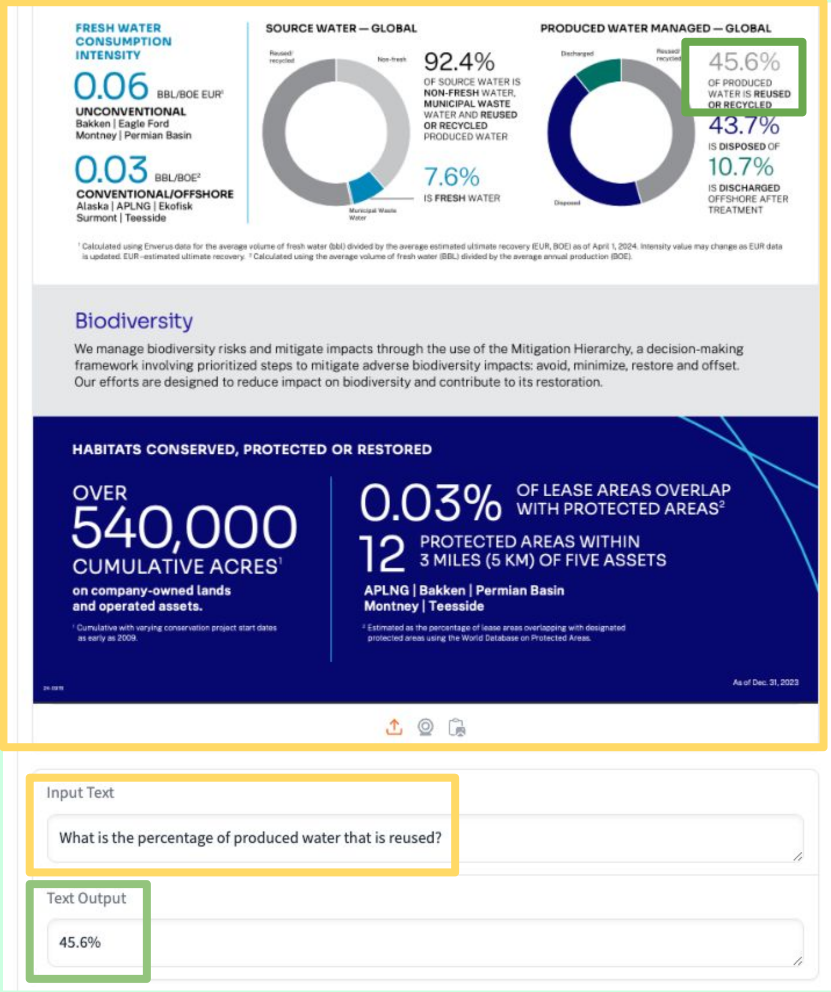
Text Input

Concatenated Tokens

Gemma

Text Output

https://huggingface.co/blog/paligemma



**FRESH WATER CONSUMPTION INTENSITY**

0.06 BBL/BOE EUR[1]

**UNCONVENTIONAL**
Bakken | Eagle Ford
Montney | Permian Basin

0.03 BBL/BOE[2]

**CONVENTIONAL/OFFSHORE**
Alaska | APLNG | Ekofisk
Surmont | Teesside

**SOURCE WATER — GLOBAL**

92.4% OF SOURCE WATER IS **NON-FRESH** WATER, **MUNICIPAL WASTE** WATER AND **REUSED OR RECYCLED** PRODUCED WATER

7.6% IS **FRESH** WATER

**PRODUCED WATER MANAGED — GLOBAL**

45.6% OF PRODUCED WATER IS **REUSED OR RECYCLED**

43.7% IS **DISPOSED** OF

10.7% IS **DISCHARGED** OFFSHORE AFTER TREATMENT

[1] Calculated using Enverus data for the average volume of fresh water (bbl) divided by the average estimated ultimate recovery (EUR, BOE) as of April 1, 2024. Intensity value may change as EUR data is updated. EUR–estimated ultimate recovery. [2] Calculated using the average volume of fresh water (BBL) divided by the average annual production (BOE).

## Biodiversity

We manage biodiversity risks and mitigate impacts through the use of the Mitigation Hierarchy, a decision-making framework involving prioritized steps to mitigate adverse biodiversity impacts: avoid, minimize, restore and offset. Our efforts are designed to reduce impact on biodiversity and contribute to its restoration.

**HABITATS CONSERVED, PROTECTED OR RESTORED**

OVER 540,000 CUMULATIVE ACRES[1]
**on company-owned lands and operated assets.**

[1] Cumulative with varying conservation project start dates as early as 2009.

0.03% OF LEASE AREAS OVERLAP WITH PROTECTED AREAS[2]

12 PROTECTED AREAS WITHIN 3 MILES (5 KM) OF FIVE ASSETS

**APLNG | Bakken | Permian Basin**
**Montney | Teesside**

[2] Estimated as the percentage of lease areas overlapping with designated protected areas using the World Database on Protected Areas.

As of Dec. 31, 2023

Input Text

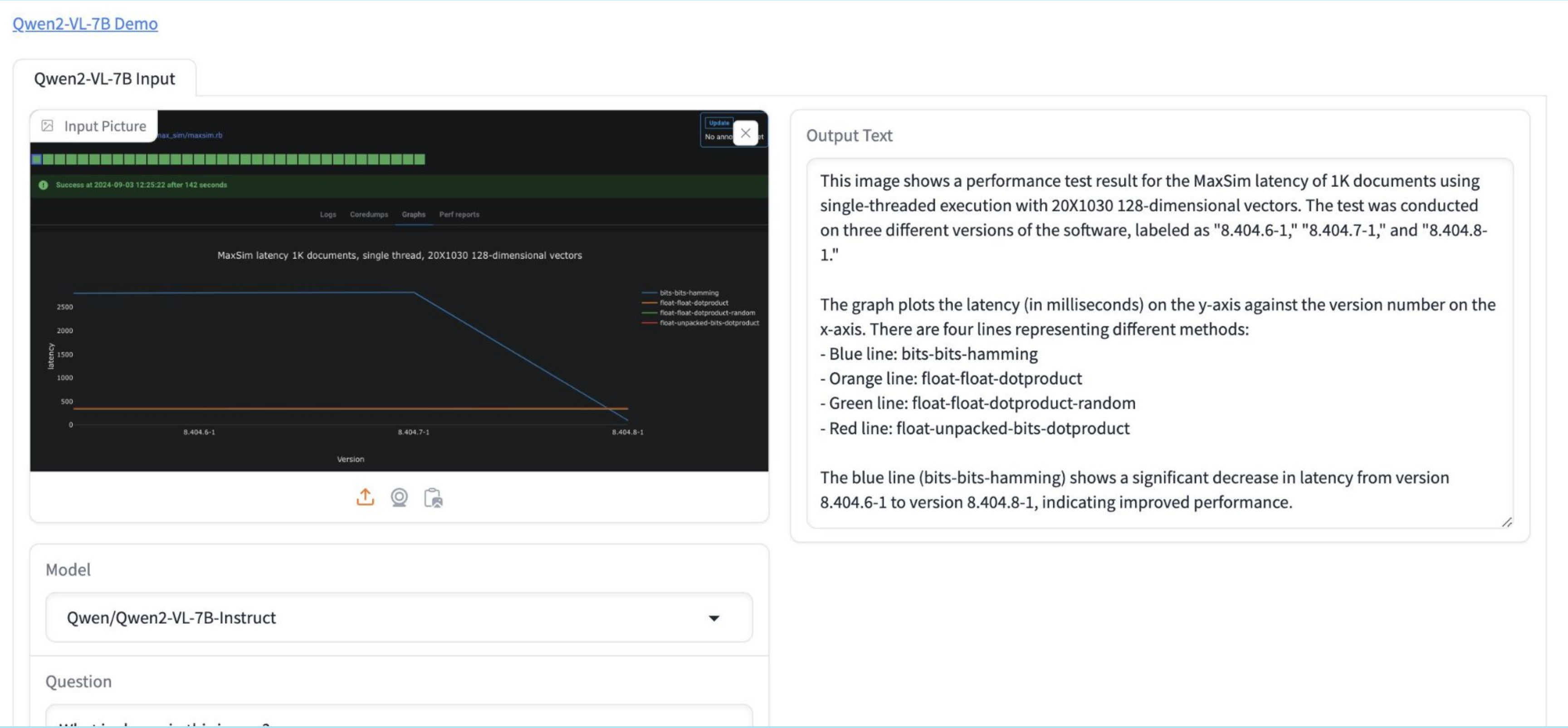
What is the percentage of produced water that is reused?

Text Output

45.6%

https://huggingface.co/spaces/big-vision/paligemma-hf

# “*What’s in the image*”



Success at 2024-09-03 12:25:22 after 142 seconds

This image shows a performance test result for the MaxSim latency of 1K documents using single-threaded execution with 20X1030 128-dimensional vectors. The test was conducted on three different versions of the software, labeled as "8.404.6-1," "8.404.7-1," and "8.404.8-1."

The graph plots the latency (in milliseconds) on the y-axis against the version number on the x-axis. There are four lines representing different methods:

- Blue line: bits-bits-hamming
- Orange line: float-float-dotproduct
- Green line: float-float-dotproduct-random
- Red line: float-unpacked-bits-dotproduct

The blue line (bits-bits-hamming) shows a significant decrease in latency from version 8.404.6-1 to version 8.404.8-1, indicating improved performance.

Model

Qwen/Qwen2-VL-7B-Instruct

# Agenda

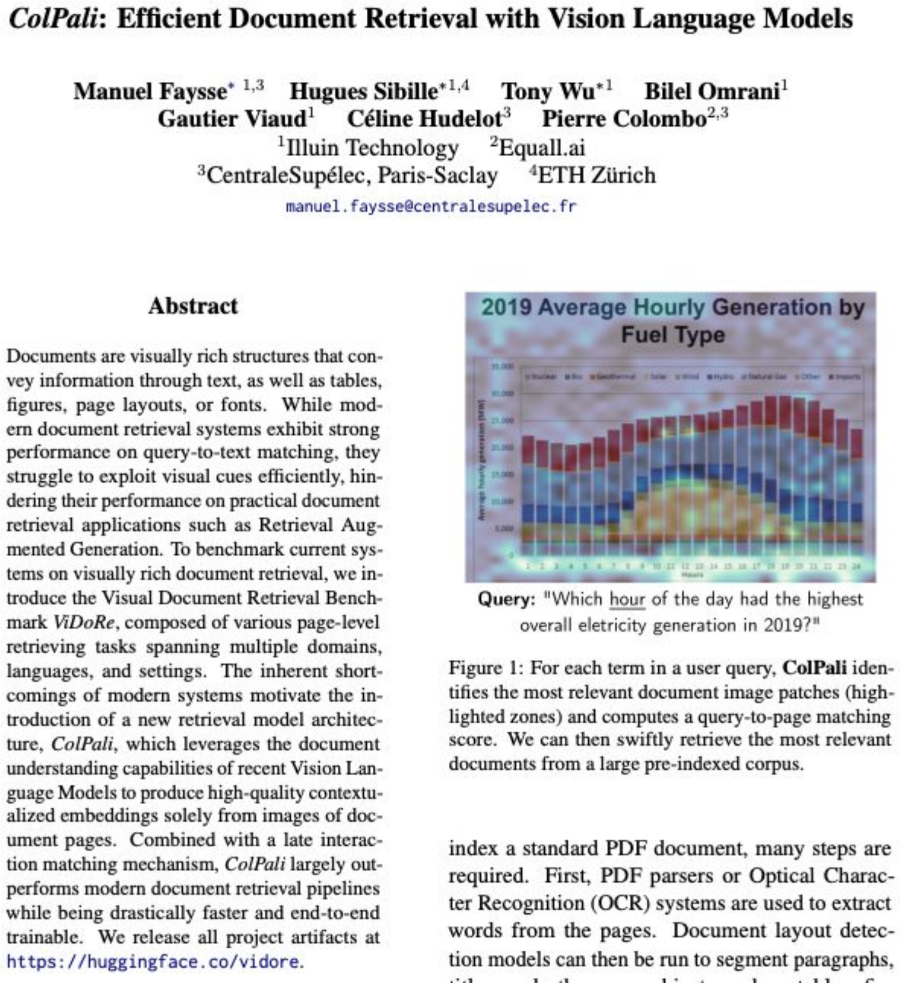
1. ~~Introduction (where we are now)~~
2. ~~Indexing & Searching PDFs~~
3. ~~Vision Language Models (VLMs)~~
4. **ColPali - Using VLM for retrieval**
5. Scaling ColPali to large collections
6. RAG with VLMs
7. QA & Resources

# ColPali: Use VLM document understanding capabilities for document retrieval

IR benchmarks measures clean pre-processed texts

**In practical industrial settings we don't have the luxury of pre-processed datasets.**

Performance bottleneck usually not in embedding model performance or chunking, but in the data ingestion pipeline



## *ColPali*: Efficient Document Retrieval with Vision Language Models

**Manuel Faysse***[1,3] **Hugues Sibille***[1,4] **Tony Wu***[1] **Bilel Omrani**[1]
**Gautier Viaud**[1] **Céline Hudelot**[3] **Pierre Colombo**[2,3]
[1]Illuin Technology [2]Equall.ai
[3]CentraleSupélec, Paris-Saclay [4]ETH Zürich
manuel.faysse@centralesupelec.fr

### Abstract

Documents are visually rich structures that convey information through text, as well as tables, figures, page layouts, or fonts. While modern document retrieval systems exhibit strong performance on query-to-text matching, they struggle to exploit visual cues efficiently, hindering their performance on practical document retrieval applications such as Retrieval Augmented Generation. To benchmark current systems on visually rich document retrieval, we introduce the Visual Document Retrieval Benchmark *ViDoRe*, composed of various page-level retrieving tasks spanning multiple domains, languages, and settings. The inherent shortcomings of modern systems motivate the introduction of a new retrieval model architecture, *ColPali*, which leverages the document understanding capabilities of recent Vision Language Models to produce high-quality contextualized embeddings solely from images of document pages. Combined with a late interaction matching mechanism, *ColPali* largely outperforms modern document retrieval pipelines while being drastically faster and end-to-end trainable. We release all project artifacts at https://huggingface.co/vidore.

2019 Average Hourly Generation by Fuel Type

**Query:** "Which hour of the day had the highest overall eletricity generation in 2019?"

Figure 1: For each term in a user query, **ColPali** identifies the most relevant document image patches (highlighted zones) and computes a query-to-page matching score. We can then swiftly retrieve the most relevant documents from a large pre-indexed corpus.

index a standard PDF document, many steps are required. First, PDF parsers or Optical Character Recognition (OCR) systems are used to extract words from the pages. Document layout detection models can then be run to segment paragraphs,

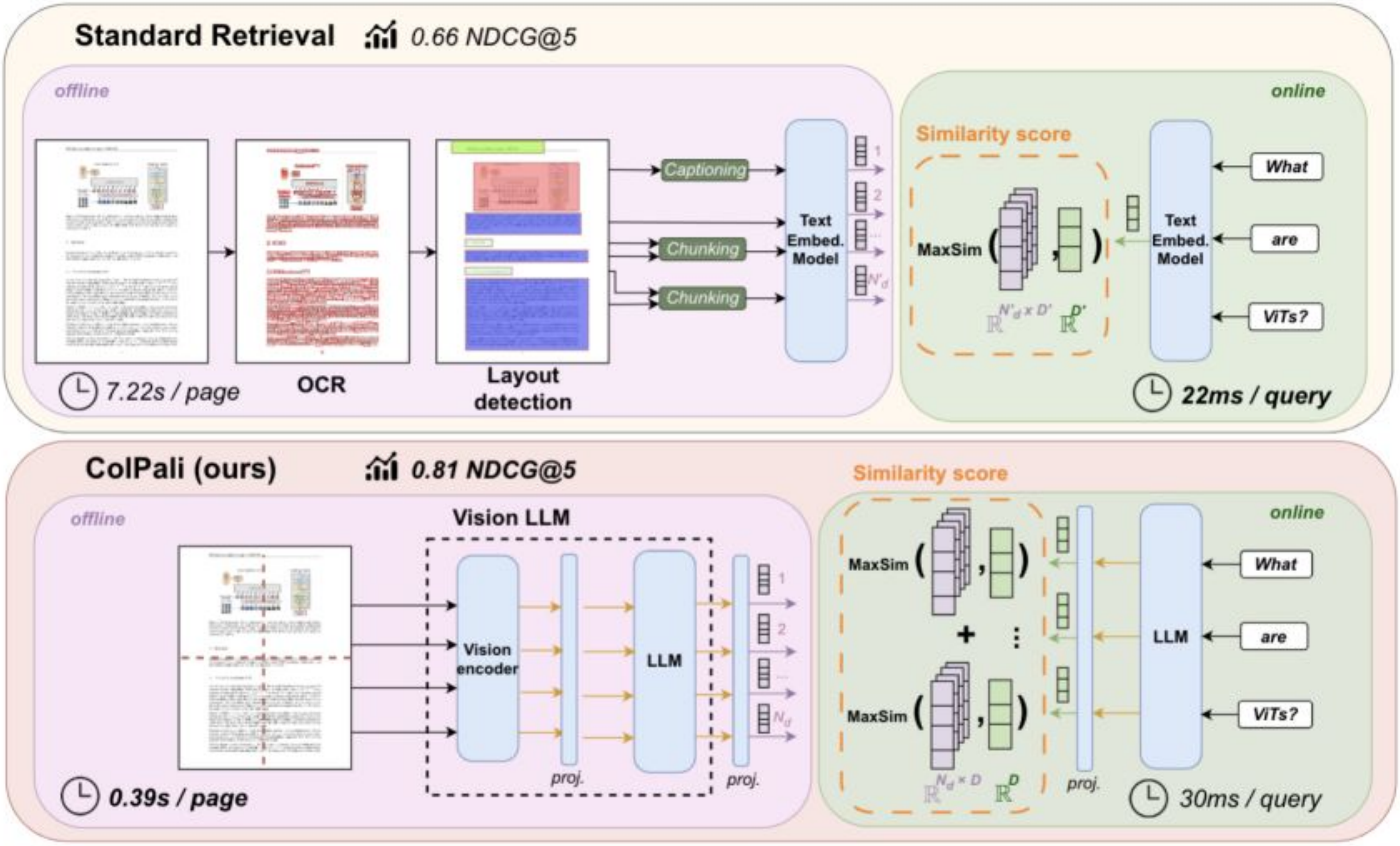


Figure 2: *ColPali* simplifies document retrieval w.r.t. standard retrieval methods while achieving stronger performances with better latencies. Latencies and results are detailed in section 5 and subsection B.5.

# ColPali (Gemma)

ColPali is short for *Contextualized Late Interaction over PaliGemma* and builds on two concepts:

- **Contextualized Embeddings from VLM**
  - **ColPali generates contextualized multi-vector embeddings directly from the screenshot of a page or text query using the VLM as the backbone**
- **Interaction between text query vectors and screenshot vectors at scoring time**

**A bi-encoder architecture, enables offline indexing - but with multi-vectors per page**

# ColPali offers

✓ **Match without text extraction**
Do not have to map complex formats to the text domain

✓ **Avoid OCR**
Use the VLM OCR capabilities

✓ **No Layout Detection**
Layout is encoded by the VLM

✓ **Engineering Simplicity**
Reduced document processing pipeline complexity

✓ **Better relevance**
Outcompete traditional extract methods

✓ **A future direction for document retrieval**
Can you print the document, then you can index it with the ColPali approach

# ColPali page level embeddings

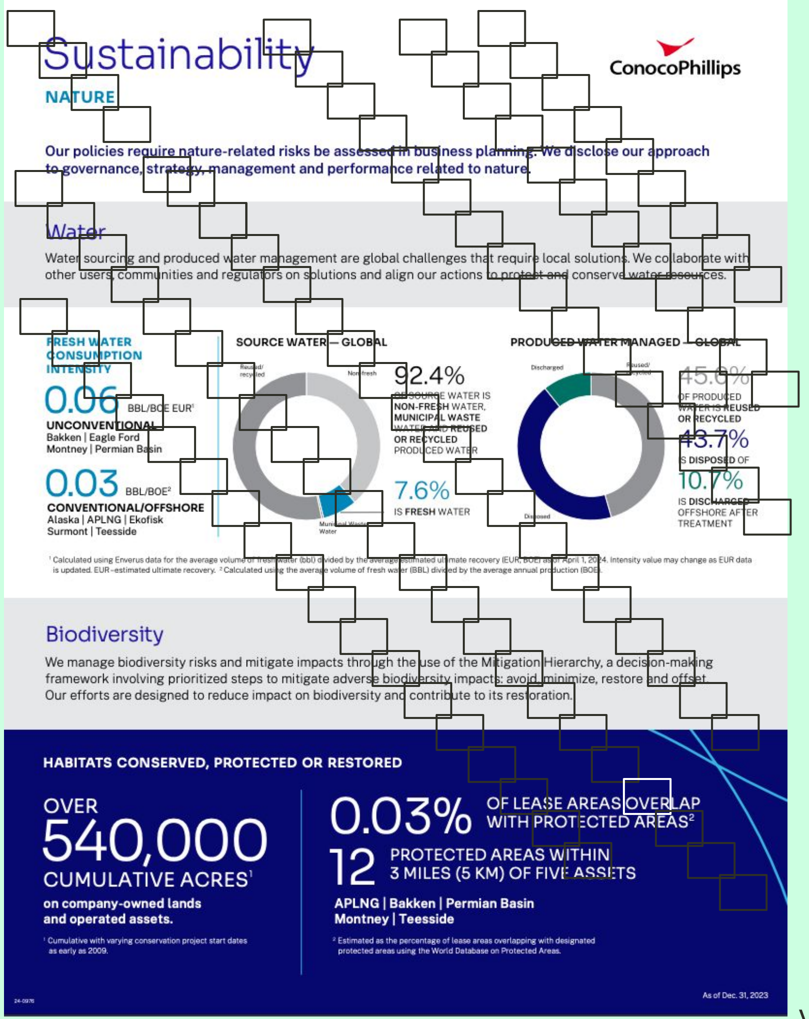
VLM “see” the image as 32x32 patches = 1024 patches.

**An image is worth 1024 words**

Each patch is represented or projected into a 128-dimensional vector space (the latent space)

6 text tokens projected to the same space = 1030 vectors per page

Tokens from text prefix “Describe the image” <img>



# Sustainability

ConocoPhillips

## NATURE

Our policies require nature-related risks be assessed in business planning. We disclose our approach to governance, strategy, management and performance related to nature.

## Water

Water sourcing and produced water management are global challenges that require local solutions. We collaborate with other users, communities and regulators on solutions and align our actions to protect and conserve water resources.

**FRESH WATER CONSUMPTION INTENSITY**

0.06 BBL/BOE EUR[1]

**UNCONVENTIONAL**
Bakken | Eagle Ford
Montney | Permian Basin

0.03 BBL/BOE[2]

**CONVENTIONAL/OFFSHORE**
Alaska | APLNG | Ekofisk
Surmont | Teesside

**SOURCE WATER — GLOBAL**

92.4% OF SOURCE WATER IS **NON-FRESH** WATER, **MUNICIPAL WASTE WATER AND REUSED OR RECYCLED** PRODUCED WATER

7.6% IS **FRESH** WATER

**PRODUCED WATER MANAGED — GLOBAL**

45.6% **OF PRODUCED WATER IS REUSED OR RECYCLED**

43.7% **IS DISPOSED** OF

10.7% **IS DISCHARGED** OFFSHORE AFTER TREATMENT

[1] Calculated using Enverus data for the average volume of fresh water (bbl) divided by the average estimated ultimate recovery (EUR, BOE) as of April 1, 2024. Intensity value may change as EUR data is updated. EUR–estimated ultimate recovery. [2] Calculated using the average volume of fresh water (BBL) divided by the average annual production (BOE).

## Biodiversity

We manage biodiversity risks and mitigate impacts through the use of the Mitigation Hierarchy, a decision-making framework involving prioritized steps to mitigate adverse biodiversity impacts: avoid, minimize, restore and offset. Our efforts are designed to reduce impact on biodiversity and contribute to its restoration.

**HABITATS CONSERVED, PROTECTED OR RESTORED**

OVER 540,000 CUMULATIVE ACRES[1]

**on company-owned lands and operated assets.**

[1] Cumulative with varying conservation project start dates as early as 2009.

0.03% OF LEASE AREAS OVERLAP WITH PROTECTED AREAS[2]

12 PROTECTED AREAS WITHIN 3 MILES (5 KM) OF FIVE ASSETS

**APLNG | Bakken | Permian Basin**
**Montney | Teesside**

[2] Estimated as the percentage of lease areas overlapping with designated protected areas using the World Database on Protected Areas.

24-0976

As of Dec. 31, 2023

# ColPali page level embeddings

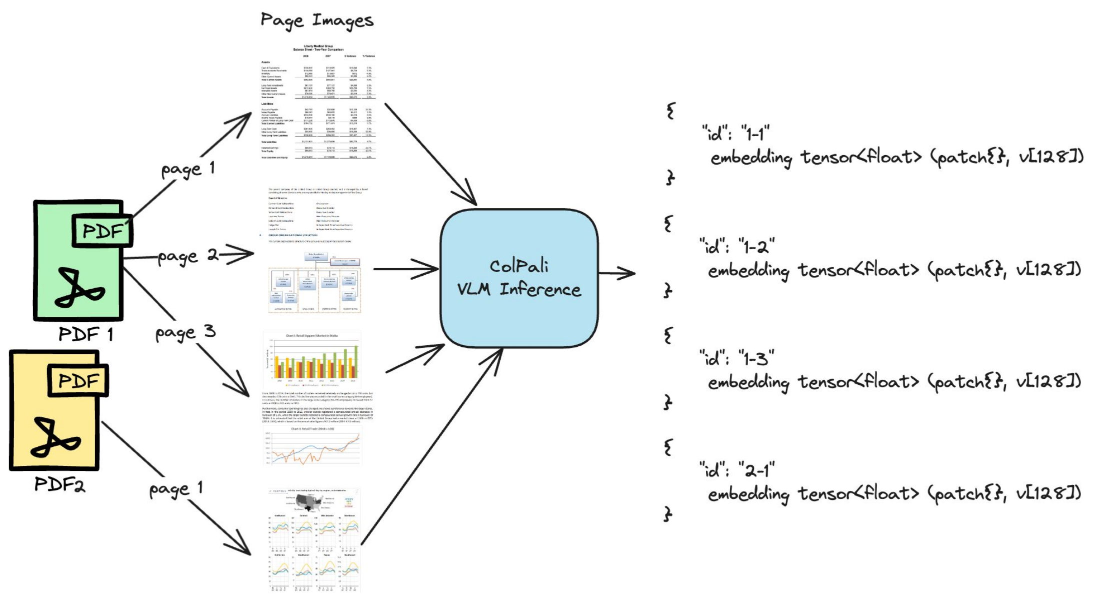
One PDF - Multiple pages

One page represented as a tensor

Vespa tensor definition

*tensor<float>(patch{}, v[128])*

A map of vectors (patch is the key, the vector the value)



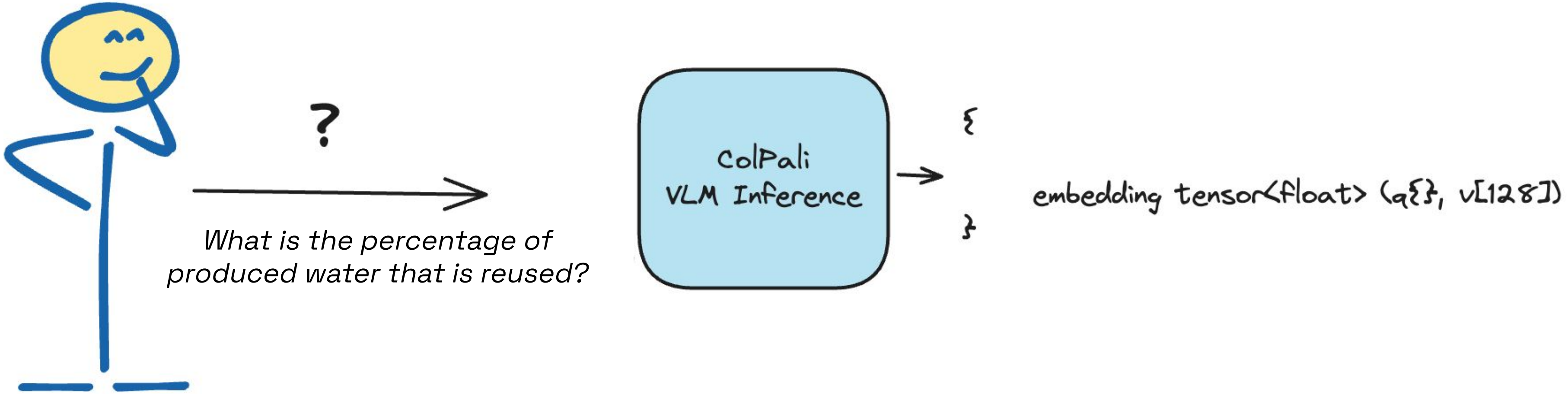
```
{
  "id": "1-1"
   embedding tensor<float> (patch{}, v[128])
}
{
  "id": "1-2"
   embedding tensor<float> (patch{}, v[128])
}
{
  "id": "1-3"
   embedding tensor<float> (patch{}, v[128])
}
{
  "id": "2-1"
   embedding tensor<float> (patch{}, v[128])
}
```

# ColPali text query embeddings

One 128-dimensional vector per user input query *token*

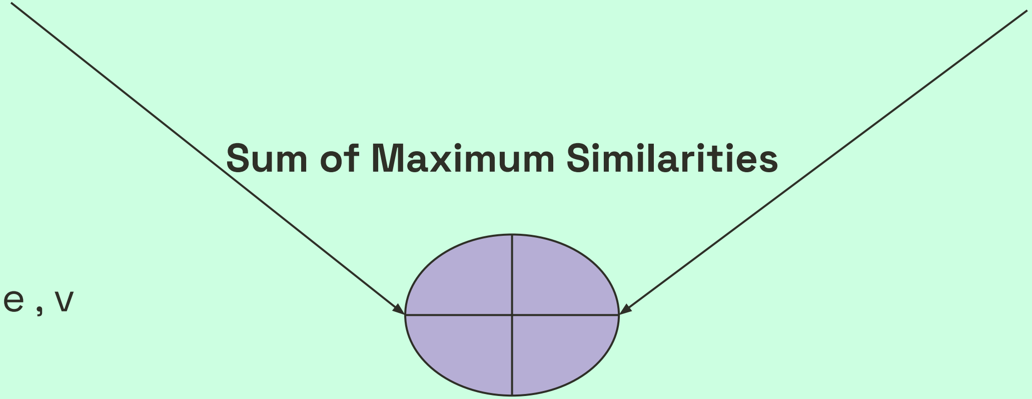
Fixed prepend and query expansion tokens

*tensor<float>(q{},v[128])*



*What is the percentage of produced water that is reused?*

# ColPali Relevance Scoring
# score(query, page):



*tensor<float>(q{}, v[128])*

*tensor<float>(p{},v[128])*

**Sum of Maximum Similarities**

```
sum(
     reduce(
       sum(
         query * page , v
       ),
       max,
       p
     ),
   q
 )
```

***Contextualized Late-Interaction over Embeddings from PaliGemma***

# (Sum) of MaxSim

Similarity matrix *|q| X |p|* - where similarity is the dot product

Find maximum patch similarity per query token. Toy example with 2-dimensional **v** and 4 image patches.

Query token vectors

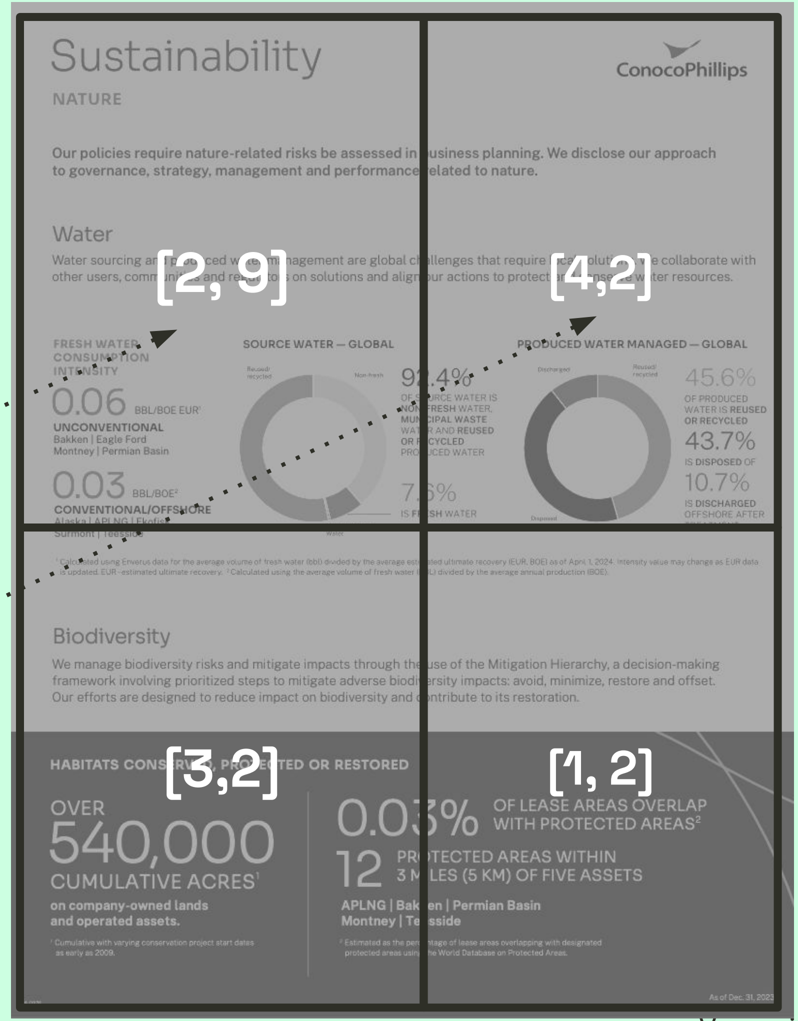
water [1, 4]

max( [**38**, 12, 11, 9]) = **38**

recycle [8,2]

max([34, **36**, 28,12]) = **36**

**score(query,page) = 38 + 36 = 74**



Sustainability

ConocoPhillips

NATURE

Our policies require nature-related risks be assessed in business planning. We disclose our approach to governance, strategy, management and performance related to nature.

Water

Water sourcing and produced water management are global challenges that require local solutions. We collaborate with other users, communities and regulators on solutions and align our actions to protect and conserve water resources.

FRESH WATER CONSUMPTION INTENSITY

0.06 BBL/BOE EUR[1]

UNCONVENTIONAL
Bakken | Eagle Ford
Montney | Permian Basin

0.03 BBL/BOE[2]

CONVENTIONAL/OFFSHORE
Alaska | APLNG | Ekofisk
Surmont | Teesside

SOURCE WATER — GLOBAL

92.4% OF SOURCE WATER IS NON-FRESH WATER, MUNICIPAL WASTE WATER AND REUSED OR RECYCLED PRODUCED WATER

7.6% IS FRESH WATER

PRODUCED WATER MANAGED — GLOBAL

45.6% OF PRODUCED WATER IS REUSED OR RECYCLED

43.7% IS DISPOSED OF

10.7% IS DISCHARGED OFFSHORE AFTER

[1] Calculated using Enverus data for the average volume of fresh water (bbl) divided by the average estimated ultimate recovery (EUR, BOE) as of April 1, 2024. Intensity value may change as EUR data is updated. EUR - estimated ultimate recovery. [2] Calculated using the average volume of fresh water (BBL) divided by the average annual production (BOE).

Biodiversity

We manage biodiversity risks and mitigate impacts through the use of the Mitigation Hierarchy, a decision-making framework involving prioritized steps to mitigate adverse biodiversity impacts: avoid, minimize, restore and offset. Our efforts are designed to reduce impact on biodiversity and contribute to its restoration.

HABITATS CONSERVED, PROTECTED OR RESTORED

OVER 540,000 CUMULATIVE ACRES[1]

on company-owned lands and operated assets.

[1] Cumulative with varying conservation project start dates as early as 2009.

0.03% OF LEASE AREAS OVERLAP WITH PROTECTED AREAS[2]

12 PROTECTED AREAS WITHIN 3 MILES (5 KM) OF FIVE ASSETS

APLNG | Bakken | Permian Basin
Montney | Teesside

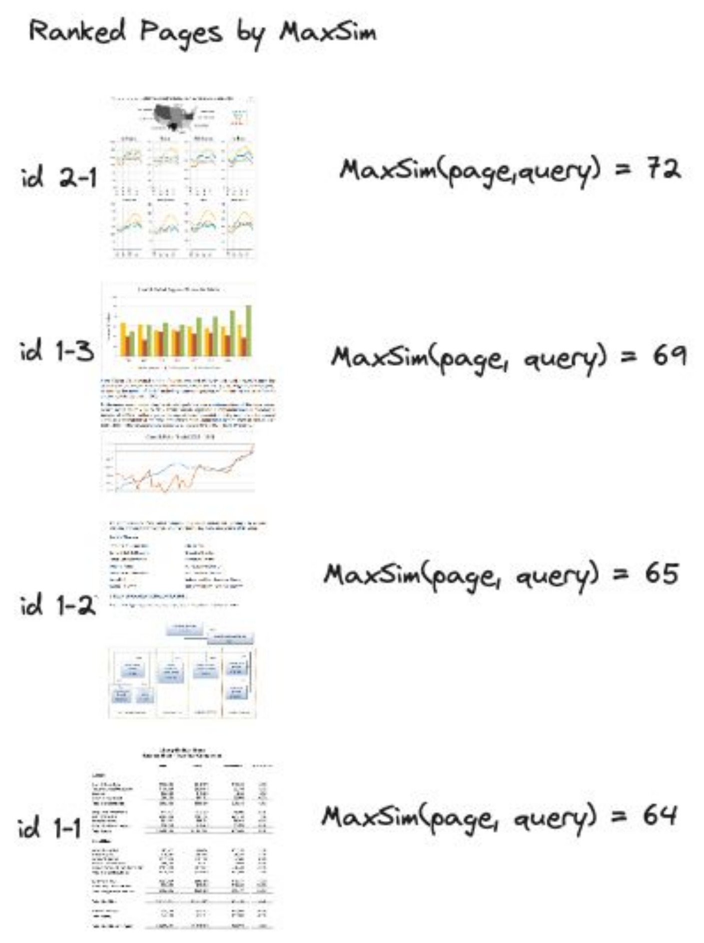
[2] Estimated as the percentage of lease areas overlapping with designated protected areas using the World Database on Protected Areas.

As of Dec. 31, 2023

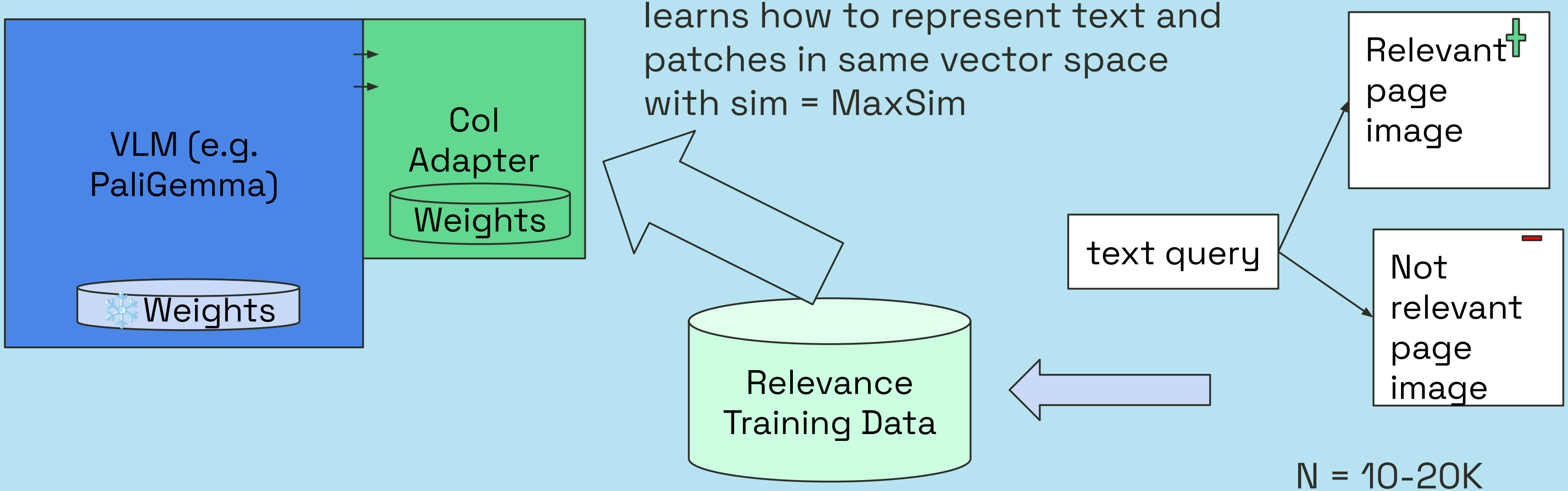
# Score each page in the collection with (Sum of) MaxSim

Similar process as with “regular” search BM25 (Lucene, Vespa, Elasticsearch)

Sort the pages by score in descending order. Score is a proxy for relevance, higher is better.



# Learning Col vectors for retrieval (representation learning)



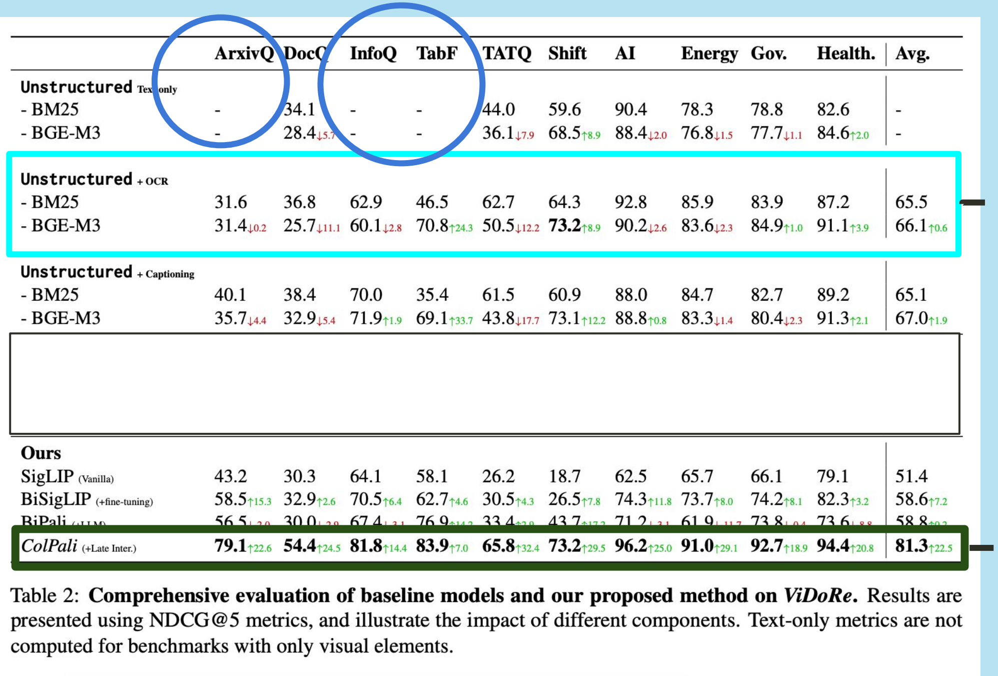
Train adapter weights which learns how to represent text and patches in same vector space with sim = MaxSim

N = 10-20K

Vespa

# How does ColPali compare with traditional methods?

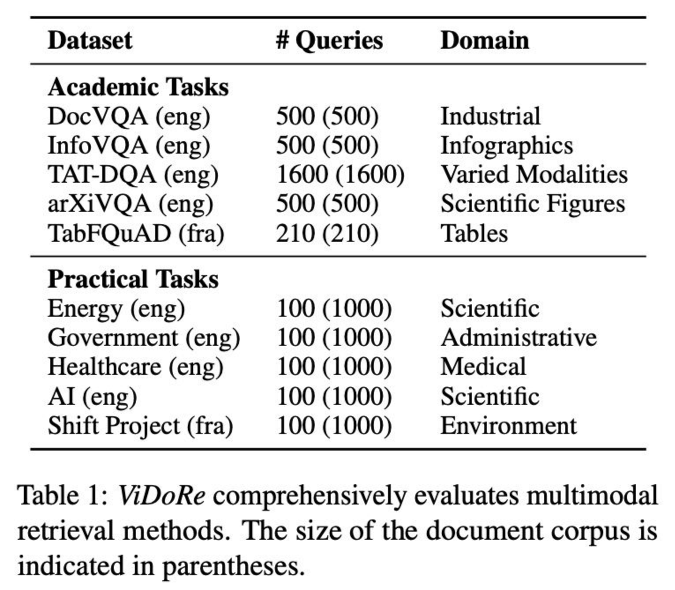
# The ViDoRe Benchmark



| | 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 |
| **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.

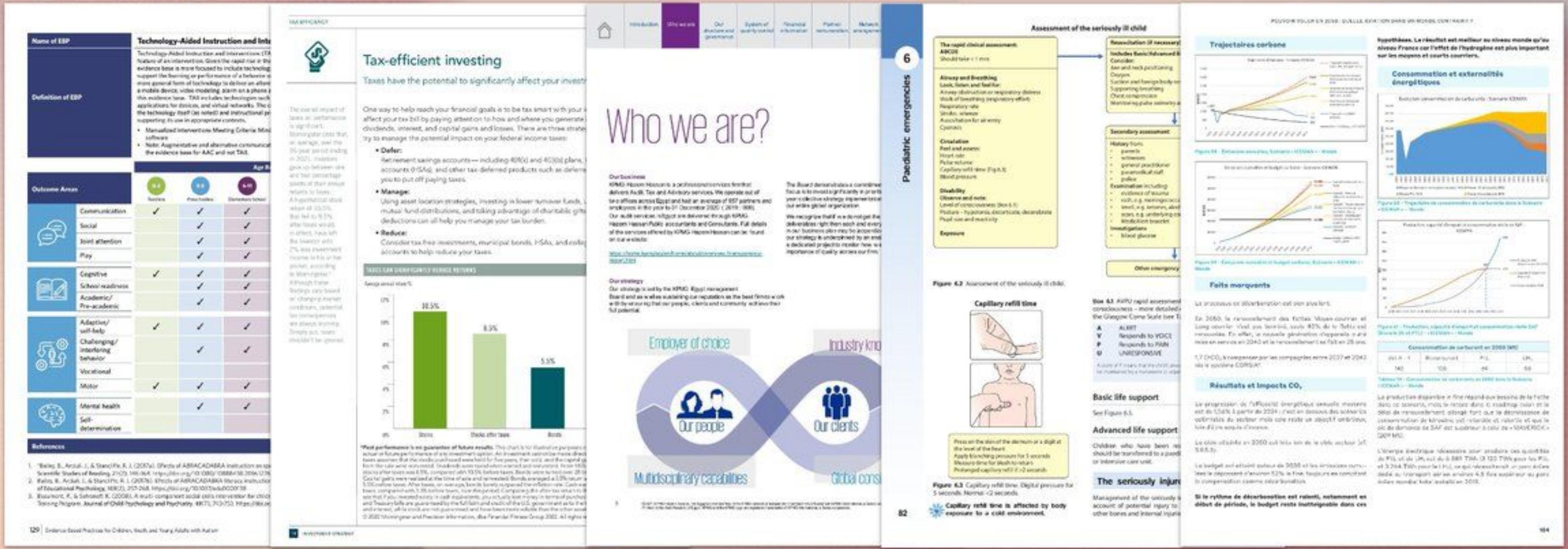
## Datasets ViDoRe



| **Dataset** | **# Queries** | **Domain** |
|---|---|---|
| **Academic Tasks** | | |
| DocVQA (eng) | 500 (500) | Industrial |
| InfoVQA (eng) | 500 (500) | Infographics |
| TAT-DQA (eng) | 1600 (1600) | Varied Modalities |
| arXiVQA (eng) | 500 (500) | Scientific Figures |
| TabFQuAD (fra) | 210 (210) | Tables |
| **Practical Tasks** | | |
| Energy (eng) | 100 (1000) | Scientific |
| Government (eng) | 100 (1000) | Administrative |
| Healthcare (eng) | 100 (1000) | Medical |
| AI (eng) | 100 (1000) | Scientific |
| Shift Project (fra) | 100 (1000) | Environment |

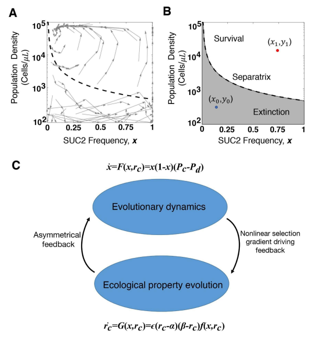
Table 1: *ViDoRe* comprehensively evaluates multimodal retrieval methods. The size of the document corpus is indicated in parentheses.

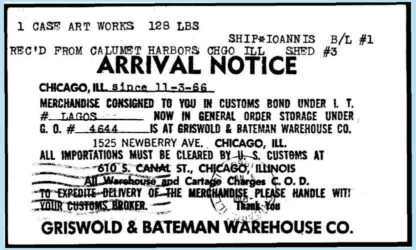
**nDCG@5 avg 0.66 => 0.813**



# Samples from the ViDoRe benchmark dataset

# Samples of PDF pages from the ViDoRe benchmark





1 CASE ART WORKS 128 LBS

SHIP*IOANNIS B/L #1

REC'D FROM CALUMET HARBORS CHGO ILL SHED #3

## ARRIVAL NOTICE

CHICAGO, ILL. since 11-3-66

MERCHANDISE CONSIGNED TO YOU IN CUSTOMS BOND UNDER I. T. # LAGOS NOW IN GENERAL ORDER STORAGE UNDER G. O. # 4644 IS AT GRISWOLD & BATEMAN WAREHOUSE CO.

1525 NEWBERRY AVE. CHICAGO, ILL.

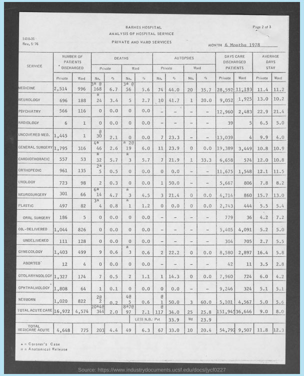
ALL IMPORTATIONS MUST BE CLEARED BY U. S. CUSTOMS AT 610 S. CANAL ST., CHICAGO, ILLINOIS

All Warehouse and Cartage Charges C. O. D.

TO EXPEDITE DELIVERY OF THE MERCHANDISE PLEASE HANDLE WITH YOUR CUSTOMS BROKER.

Thank You

GRISWOLD & BATEMAN WAREHOUSE CO.



1410-31
Rev. 5/76

BARNES HOSPITAL
ANALYSIS OF HOSPITAL SERVICE
PRIVATE AND WARD SERVICES

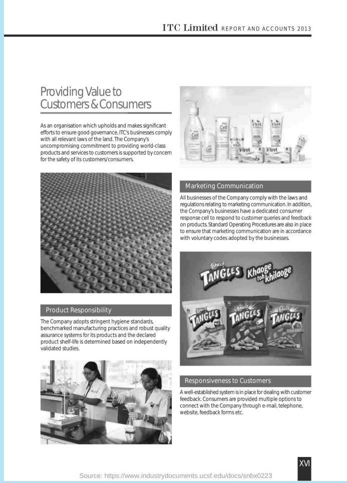
Page 2 of 3

MONTH 6 Months 1978

| SERVICE | NUMBER OF PATIENTS DISCHARGED | | DEATHS | | | | AUTOPSIES | | | | DAYS CARE DISCHARGED PATIENTS | | AVERAGE DAYS STAY | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Private | | Ward | | Private | | Ward | | | | | |
| | Private | Ward | No. | % | No. | % | No. | % | No. | % | Private | Ward | Private | Ward |
| MEDICINE | 2,514 | 996 | 3* @ 168 | 6.7 | 3* @ 56 | 5.6 | 74 | 44.0 | 20 | 35.7 | 28,592 | 11,193 | 11.4 | 11.2 |
| NEUROLOGY | 696 | 188 | * 24 | 3.4 | 5 | 2.7 | 10 | 41.7 | 1 | 20.0 | 9,052 | 1,925 | 13.0 | 10.2 |
| PSYCHIATRY | 566 | 116 | 0 | 0.0 | 0 | 0.0 | – | – | – | – | 12,960 | 2,483 | 22.9 | 21.4 |
| RADIOLOGY | 6 | 1 | 0 | 0.0 | 0 | 0.0 | – | – | – | – | 39 | 5 | 6.5 | 5.0 |
| UNCOVERED MED. | 1,445 | 1 | @ 30 | 2.1 | 0 | 0.0 | 7 | 23.3 | – | – | 13,039 | 4 | 9.9 | 4.0 |
| GENERAL SURGERY | 1,795 | 316 | 4* 46 | 2.6 | * 2@ 19 | 6.0 | 11 | 23.9 | 0 | 0.0 | 19,389 | 3,449 | 10.8 | 10.9 |
| CARDIOTHORACIC | 557 | 53 | * 32 | 5.7 | * 3 | 5.7 | 7 | 21.9 | 1 | 33.3 | 6,658 | 574 | 12.0 | 10.8 |
| ORTHOPEDIC | 961 | 135 | 2* 5 | 0.5 | 0 | 0.0 | 0 | 0.0 | – | – | 11,675 | 1,548 | 12.1 | 11.5 |
| UROLOGY | 723 | 98 | 2 | 0.3 | 0 | 0.0 | 1 | 50.0 | – | – | 5,667 | 806 | 7.8 | 8.2 |
| NEUROSURGERY | 301 | 66 | 6* 14 | 4.7 | * 3 | 4.5 | 3 | 21.4 | 0 | 0.0 | 4,714 | 860 | 15.7 | 13.0 |
| PLASTIC | 497 | 82 | 3* 4 | 0.8 | * 1 | 1.2 | 0 | 0.0 | 0 | 0.0 | 2,743 | 444 | 5.5 | 5.4 |
| ORAL SURGERY | 186 | 5 | 0 | 0.0 | 0 | 0.0 | – | – | – | – | 779 | 36 | 4.2 | 7.2 |
| OB.-DELIVERED | 1,044 | 826 | 0 | 0.0 | 0 | 0.0 | – | – | – | – | 5,405 | 4,091 | 5.2 | 5.0 |
| UNDELIVERED | 111 | 128 | 0 | 0.0 | 0 | 0.0 | – | – | – | – | 304 | 705 | 2.7 | 5.5 |
| GYNECOLOGY | 1,403 | 499 | 9 | 0.6 | * 3 | 0.6 | 2 | 22.2 | 0 | 0.0 | 8,580 | 2,897 | 16.4 | 5.8 |
| ABORTED | 12 | 4 | 0 | 0.0 | 0 | 0.0 | – | – | – | – | 42 | 11 | 3.5 | 2.8 |
| OTOLARYNGOLOGY | 1,327 | 174 | 7 | 0.5 | 2 | 1.1 | 1 | 14.3 | 0 | 0.0 | 7,960 | 724 | 6.0 | 4.2 |
| OPHTHALMOLOGY | 1,808 | 64 | 1 | 0.1 | 0 | 0.0 | 0 | 0.0 | – | – | 9,246 | 324 | 5.1 | 5.1 |
| NEWBORN | 1,020 | 822 | 2@ 2 | 0.2 | 4@ 5 | 0.6 | @ 1 | 50.0 | 3 | 60.0 | 5,101 | 4,567 | 5.0 | 5.6 |
| TOTAL ACUTE CARE | 16,972 | 4,574 | 20*4@ 344 | 2.0 | 8*7@ 97 | 2.1 | @ 117 | 34.0 | 25 | 25.8 | 151,945 | 36,646 | 9.0 | 8.0 |
| | | | | | | LESS N.B.: Pvt | | 33.9 | Wd | 23.9 | | | | |
| TOTAL MEDICARE ACUTE | 4,648 | 775 | 203 | 4.4 | 49 | 6.3 | 67 | 33.0 | 10 | 20.4 | 54,792 | 9,507 | 11.8 | 12.3 |

* = Coroner's Case
@ = Anatomical Release

Source: https://www.industrydocuments.ucsf.edu/docs/jycf0227



**ITC Limited** REPORT AND ACCOUNTS 2013

## Providing Value to Customers & Consumers

As an organisation which upholds and makes significant efforts to ensure good governance, ITC's businesses comply with all relevant laws of the land. The Company's uncompromising commitment to providing world-class products and services to customers is supported by concern for the safety of its customers/consumers.

### Marketing Communication

All businesses of the Company comply with the laws and regulations relating to marketing communication. In addition, the Company's businesses have a dedicated consumer response cell to respond to customer queries and feedback on products. Standard Operating Procedures are also in place to ensure that marketing communication are in accordance with voluntary codes adopted by the businesses.

### Product Responsibility

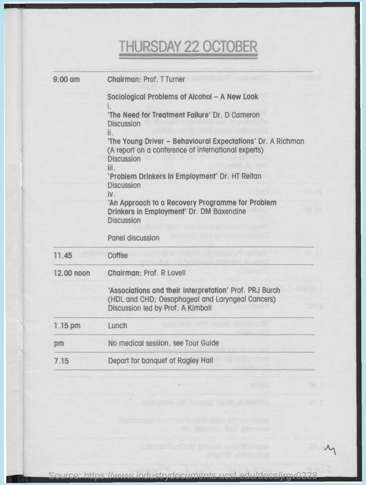
The Company adopts stringent hygiene standards, benchmarked manufacturing practices and robust quality assurance systems for its products and the declared product shelf-life is determined based on independently validated studies.

### Responsiveness to Customers

A well-established system is in place for dealing with customer feedback. Consumers are provided multiple options to connect with the Company through e-mail, telephone, website, feedback forms etc.

XVI

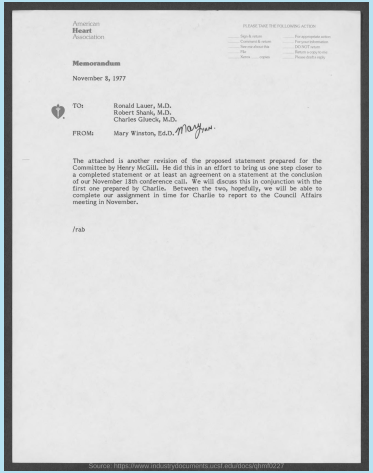
Source: https://www.industrydocuments.ucsf.edu/docs/snbx0223



## THURSDAY 22 OCTOBER

| | |
|---|---|
| 9.00 am | Chairman: Prof. T Turner<br>Sociological Problems of Alcohol – A New Look<br>i.<br>'The Need for Treatment Failure' Dr. D Cameron<br>Discussion<br>ii.<br>'The Young Driver – Behavioural Expectations' Dr. A Richman<br>(A report on a conference of international experts)<br>Discussion<br>iii.<br>'Problem Drinkers in Employment' Dr. HT Reitan<br>Discussion<br>iv.<br>'An Approach to a Recovery Programme for Problem Drinkers in Employment' Dr. DM Baxendine<br>Discussion<br>Panel discussion |
| 11.45 | Coffee |
| 12.00 noon | Chairman: Prof. R Lovell<br>'Associations and their Interpretation' Prof. PRJ Burch<br>(HDL and CHD; Oesophageal and Laryngeal Cancers)<br>Discussion led by Prof. A Kimball |
| 1.15 pm | Lunch |
| pm | No medical session, see Tour Guide |
| 7.15 | Depart for banquet at Ragley Hall |

Source: https://www.industrydocuments.ucsf.edu/docs/jrgv0228



American
**Heart**
Association

PLEASE TAKE THE FOLLOWING ACTION

____ Sign & return
____ Comment & return
____ See me about this
____ File
____ Xerox ____ copies
____ For appropriate action
____ For your information
____ DO NOT return
____ Return a copy to me
____ Please draft a reply

**Memorandum**

November 8, 1977

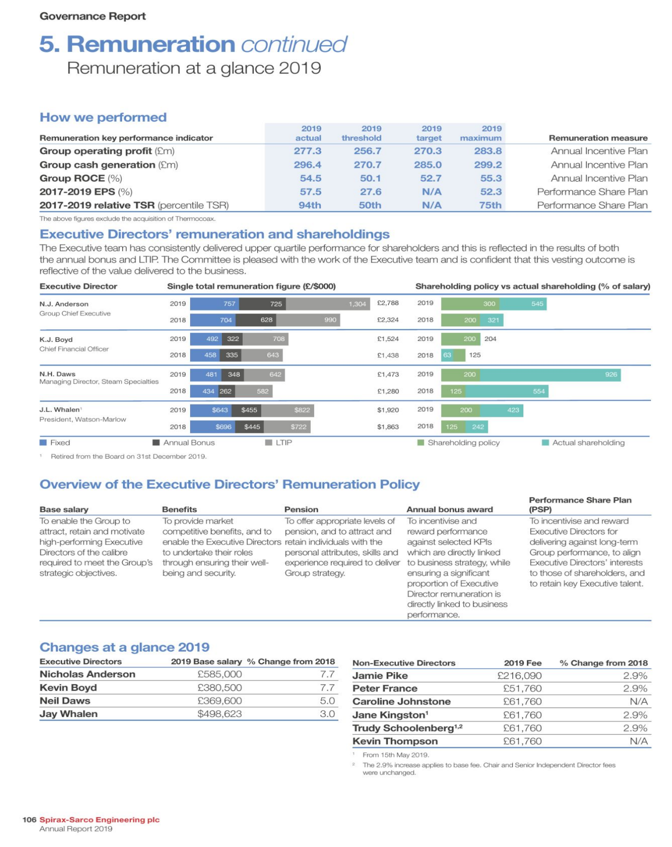
TO: Ronald Lauer, M.D.
Robert Shank, M.D.
Charles Glueck, M.D.

FROM: Mary Winston, Ed.D. Mary

The attached is another revision of the proposed statement prepared for the Committee by Henry McGill. He did this in an effort to bring us one step closer to a completed statement or at least an agreement on a statement at the conclusion of our November 18th conference call. We will discuss this in conjunction with the first one prepared by Charlie. Between the two, hopefully, we will be able to complete our assignment in time for Charlie to report to the Council Affairs meeting in November.

/rab

Source: https://www.industrydocuments.ucsf.edu/docs/qhmf0227



Governance Report

## 5. Remuneration *continued*

### Remuneration at a glance 2019

### How we performed

| Remuneration key performance indicator | 2019 actual | 2019 threshold | 2019 target | 2019 maximum | Remuneration measure |
|---|---|---|---|---|---|
| **Group operating profit** (£m) | 277.3 | 256.7 | 270.3 | 283.8 | Annual Incentive Plan |
| **Group cash generation** (£m) | 296.4 | 270.7 | 285.0 | 299.2 | Annual Incentive Plan |
| **Group ROCE** (%) | 54.5 | 50.1 | 52.7 | 55.3 | Annual Incentive Plan |
| **2017-2019 EPS** (%) | 57.5 | 27.6 | N/A | 52.3 | Performance Share Plan |
| **2017-2019 relative TSR** (percentile TSR) | 94th | 50th | N/A | 75th | Performance Share Plan |

The above figures exclude the acquisition of Thermocoax.

### Executive Directors' remuneration and shareholdings

The Executive team has consistently delivered upper quartile performance for shareholders and this is reflected in the results of both the annual bonus and LTIP. The Committee is pleased with the work of the Executive team and is confident that this vesting outcome is reflective of the value delivered to the business.

| Executive Director | Year | Fixed | Annual Bonus | LTIP | Single total remuneration figure (£/$000) | Year | Shareholding policy | Actual shareholding |
|---|---|---|---|---|---|---|---|---|
| **N.J. Anderson**<br>Group Chief Executive | 2019 | 757 | 725 | 1,304 | £2,788 | 2019 | 300 | 545 |
| | 2018 | 704 | 628 | 990 | £2,324 | 2018 | 200 | 321 |
| **K.J. Boyd**<br>Chief Financial Officer | 2019 | 492 | 322 | 708 | £1,524 | 2019 | 200 | 204 |
| | 2018 | 458 | 335 | 643 | £1,438 | 2018 | 125 | 63 |
| **N.H. Daws**<br>Managing Director, Steam Specialties | 2019 | 481 | 348 | 642 | £1,473 | 2019 | 200 | 926 |
| | 2018 | 434 | 262 | 582 | £1,280 | 2018 | 125 | 554 |
| **J.L. Whalen**[1]<br>President, Watson-Marlow | 2019 | $643 | $455 | $822 | $1,920 | 2019 | 200 | 423 |
| | 2018 | $696 | $445 | $722 | $1,863 | 2018 | 125 | 242 |

[1] Retired from the Board on 31st December 2019.

### Overview of the Executive Directors' Remuneration Policy

| Base salary | Benefits | Pension | Annual bonus award | Performance Share Plan (PSP) |
|---|---|---|---|---|
| To enable the Group to attract, retain and motivate high-performing Executive Directors of the calibre required to meet the Group's strategic objectives. | To provide market competitive benefits, and to enable the Executive Directors to undertake their roles through ensuring their well-being and security. | To offer appropriate levels of pension, and to attract and retain individuals with the personal attributes, skills and experience required to deliver Group strategy. | To incentivise and reward performance against selected KPIs which are directly linked to business strategy, while ensuring a significant proportion of Executive Director remuneration is directly linked to business performance. | To incentivise and reward Executive Directors for delivering against long-term Group performance, to align Executive Directors' interests to those of shareholders, and to retain key Executive talent. |

### Changes at a glance 2019

| Executive Directors | 2019 Base salary | % Change from 2018 |
|---|---|---|
| **Nicholas Anderson** | £585,000 | 7.7 |
| **Kevin Boyd** | £380,500 | 7.7 |
| **Neil Daws** | £369,600 | 5.0 |
| **Jay Whalen** | $498,623 | 3.0 |

| Non-Executive Directors | 2019 Fee | % Change from 2018 |
|---|---|---|
| **Jamie Pike** | £216,090 | 2.9% |
| **Peter France** | £51,760 | 2.9% |
| **Caroline Johnstone** | £61,760 | N/A |
| **Jane Kingston**[1] | £61,760 | 2.9% |
| **Trudy Schoolenberg**[1,2] | £61,760 | 2.9% |
| **Kevin Thompson** | £61,760 | N/A |

[1] From 15th May 2019.
[2] The 2.9% increase applies to base fee. Chair and Senior Independent Director fees were unchanged.

106 Spirax-Sarco Engineering plc
Annual Report 2019

Vespa.ai

# Agenda

1. ~~Introduction (where we are now)~~
2. ~~Indexing & Searching PDFs~~
3. ~~Vision Language Models (VLMs)~~
4. ~~ColPali - Using VLM for retrieval~~
5. **Scaling ColPali to large collections**
6. RAG with VLMs
7. QA & Resources

# “But, Does It Scale?”

“That’s a lot of vectors!”

Reply Guys
20.08.2024

# Scaling ColPali MaxSim

Compute scales with number of dot products **D** x [**q** x **p** x **v**]

Storage scales with **D** x [**p** x **v** ]

Note: MaxSim only, inference of model is scaled independently

- **q** is the number of query tokens, including the mask and instruction tokens
- **p** is the number of image patches
- **v** is the vector dimensionality
- **D** is the number of pages scored by MaxSim

# Scaling ColPali MaxSim

- Reduce # patch vectors (reduces |p|) - clustering, remove redundant

- Reduce precision of v from float to bit - 32x

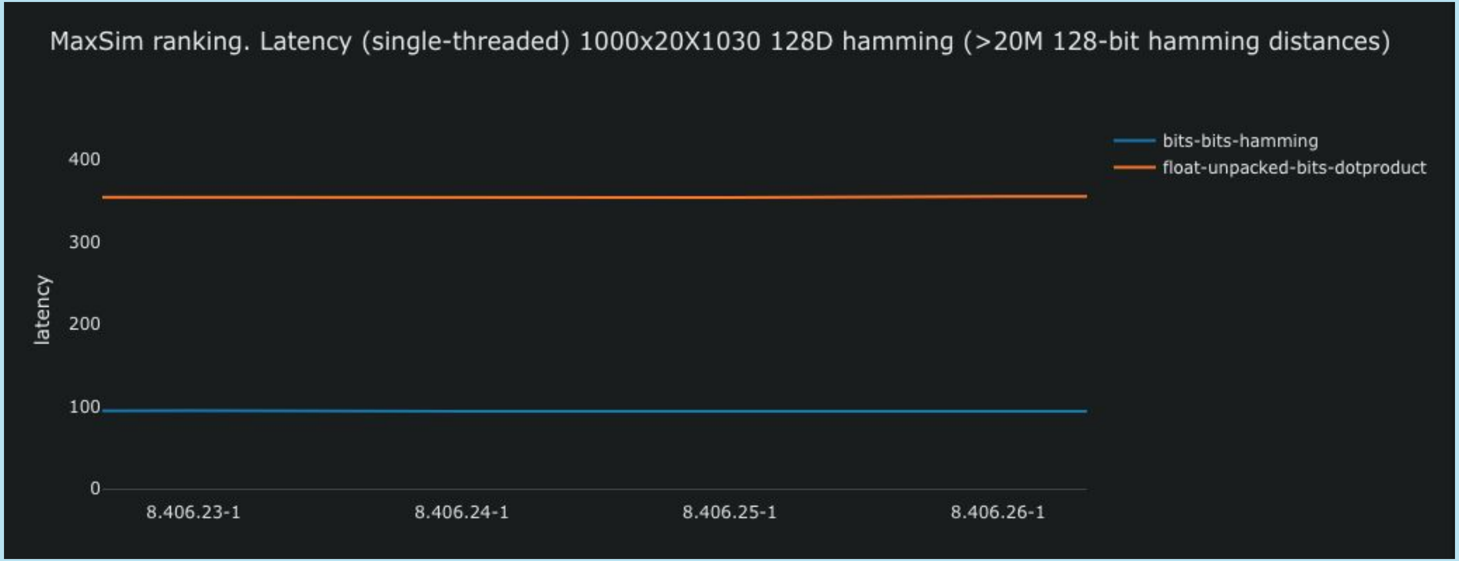
- Replace float dot products with inverted hamming (4x reduction in compute)

- **q** is the number of query tokens, including the mask and instruction tokens
- **p** is the number of image patches
- **v** is the vector dimensionality
- **D** is the number of pages scored by MaxSim

# ColPali vectors

**32x reduction storage**
**4x compute reduction**

Binary Quantization ( > 0 )
Normalized vectors (unit length). Values close to 0
Replace float dot product with inverted hamming distance (correlates)



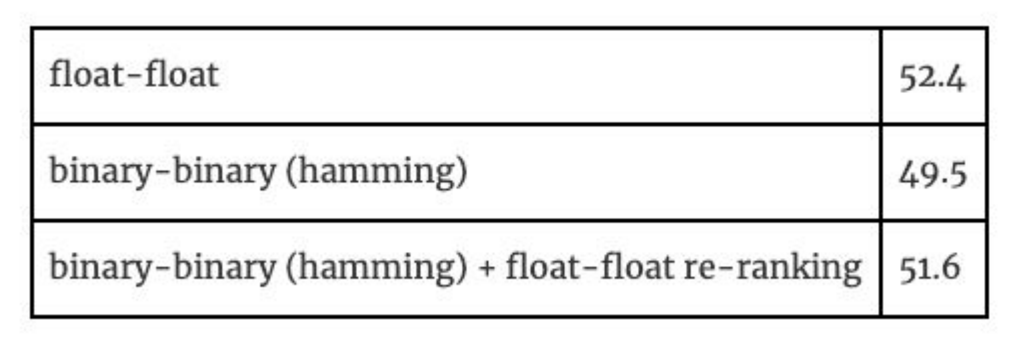
# Scaling ColPali MaxSim + Retrieval

Phased retrieval & ranking

Search for close neighbors of all k query token vectors - compatible with HNSW indexing using hamming distance metric

Step 2 - re-rank using MaxSim

nDCG@5 DocVQA

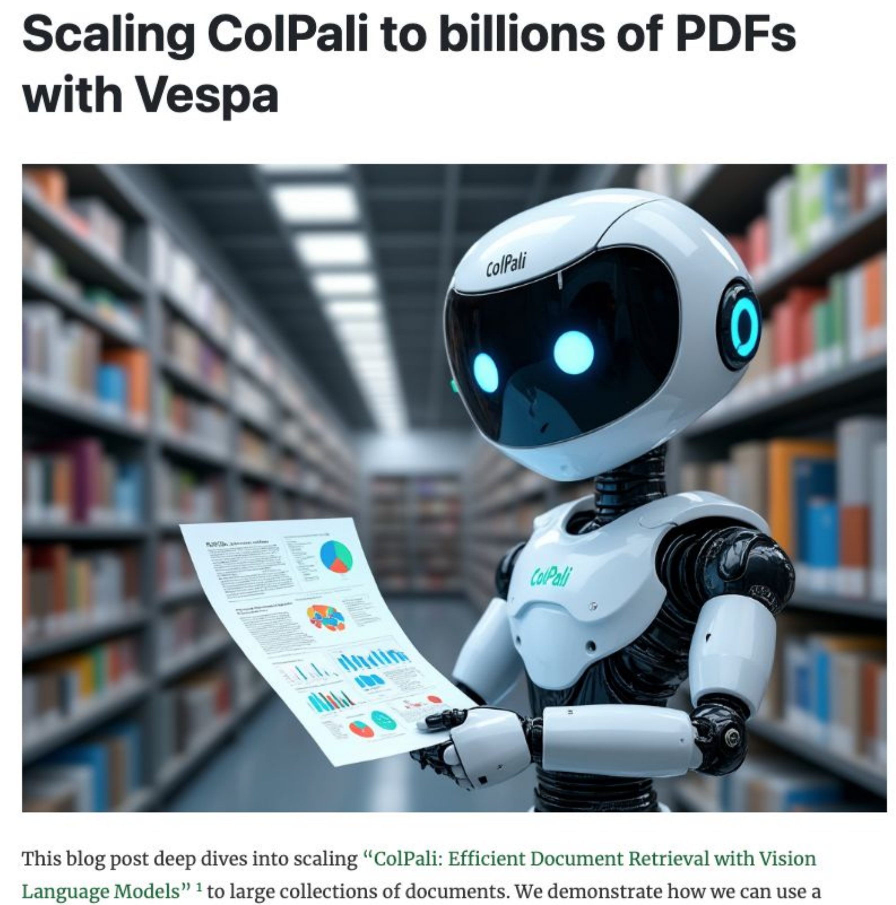


| | |
|---|---|
| float-float | 52.4 |
| binary-binary (hamming) | 49.5 |
| binary-binary (hamming) + float-float re-ranking | 51.6 |

# Scaling ColPali TLDR;

- Reduce precision (float to bit)
- Binary quantization (BQ)
- Hamming instead of dot product
- Multi-vector HNSW indexing
- Phased retrieval & ranking

[https://blog.vespa.ai/scaling-colpali-to-billions/](https://blog.vespa.ai/scaling-colpali-to-billions/)



## Scaling ColPali to billions of PDFs with Vespa

This blog post deep dives into scaling "ColPali: Efficient Document Retrieval with Vision Language Models" [1] to large collections of documents. We demonstrate how we can use a

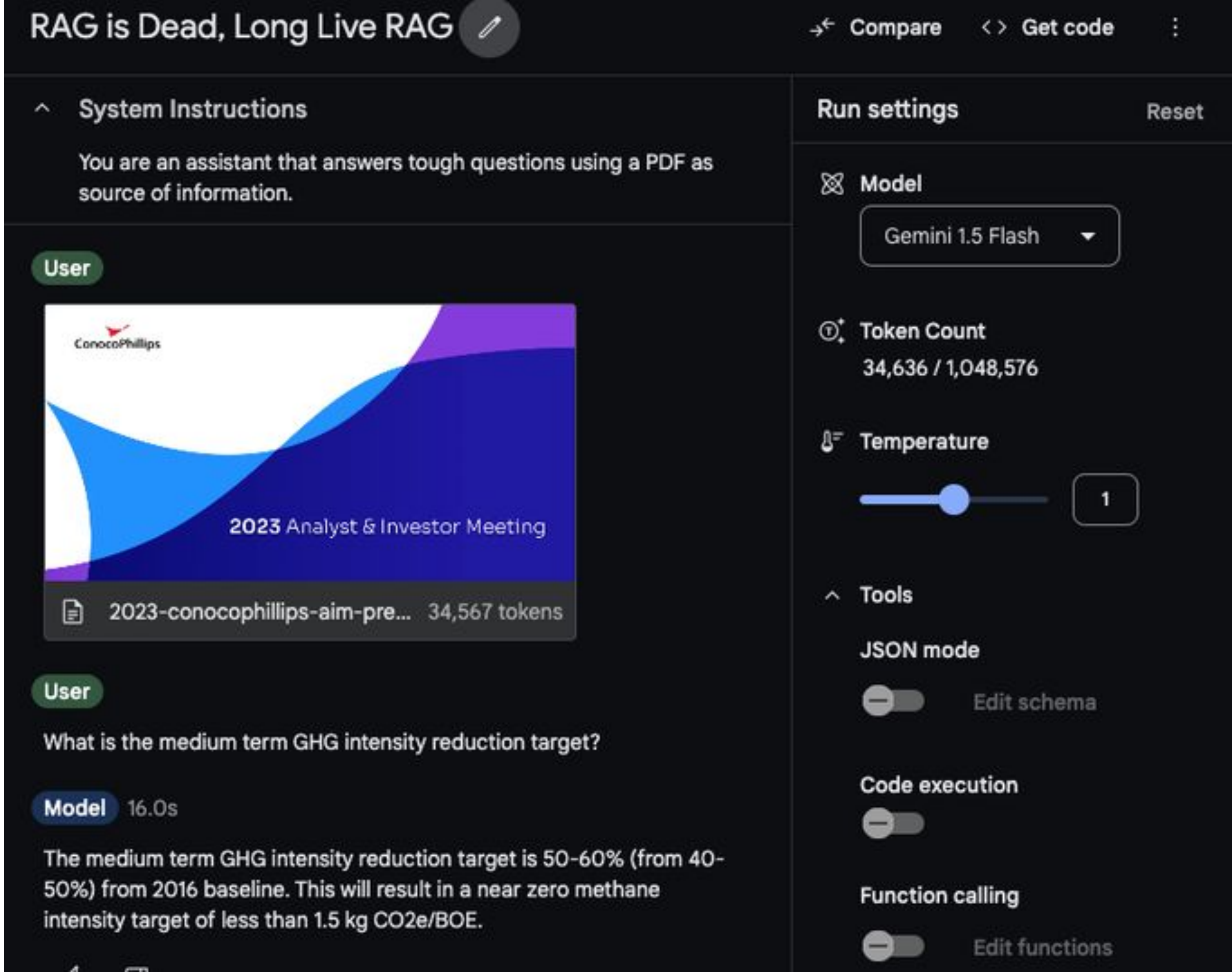
# Agenda

1. ~~Introduction (where we are now)~~
2. ~~Indexing & Searching PDFs~~
3. ~~Vision Language Models (VLMs)~~
4. ~~ColPali - Using VLM for retrieval~~
5. ~~Scaling ColPali to large collections~~
6. **RAG with VLMs**
7. QA & Resources

# RAG with ColPali

ColPali is the first stage retriever

For the generative step you need a VLM for question answering based on the retrieved context



RAG is Dead, Long Live RAG

Compare
Get code
Run settings
Reset

System Instructions

You are an assistant that answers tough questions using a PDF as source of information.

**User**

ConocoPhillips
2023 Analyst & Investor Meeting
2023-conocophillips-aim-pre... 34,567 tokens

**User**

What is the medium term GHG intensity reduction target?

**Model** 16.0s

The medium term GHG intensity reduction target is 50-60% (from 40-50%) from 2016 baseline. This will result in a near zero methane intensity target of less than 1.5 kg CO2e/BOE.

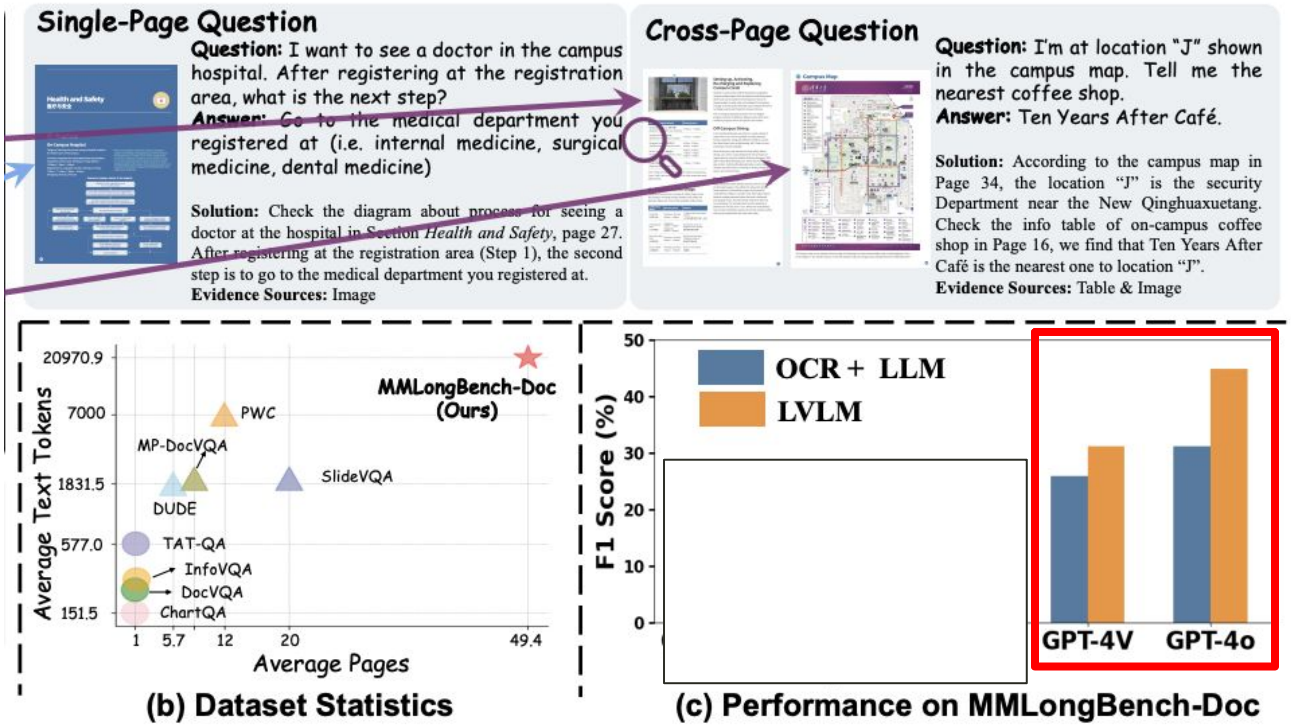
Model: Gemini 1.5 Flash

Token Count: 34,636 / 1,048,576

Temperature: 1

# Frontier VLMs

VLMs are better at question answering when using image data than text + OCR.



**Single-Page Question**

**Question:** I want to see a doctor in the campus hospital. After registering at the registration area, what is the next step?

**Answer:** Go to the medical department you registered at (i.e. internal medicine, surgical medicine, dental medicine)

**Solution:** Check the diagram about process for seeing a doctor at the hospital in Section *Health and Safety*, page 27. After registering at the registration area (Step 1), the second step is to go to the medical department you registered at.

**Evidence Sources:** Image

**Cross-Page Question**

**Question:** I'm at location "J" shown in the campus map. Tell me the nearest coffee shop.

**Answer:** Ten Years After Café.

**Solution:** According to the campus map in Page 34, the location "J" is the security Department near the New Qinghuaxuetang. Check the info table of on-campus coffee shop in Page 16, we find that Ten Years After Café is the nearest one to location "J".

**Evidence Sources:** Table & Image

(b) Dataset Statistics

(c) Performance on MMLongBench-Doc

From **MMLongBench-Doc: Benchmarking Long-context Document Understanding with Visualizations**

# Agenda

1. ~~Introduction (where we are now)~~
2. ~~Indexing & Searching PDFs~~
3. ~~Vision Language Models (VLMs)~~
4. ~~ColPali - Using VLM for retrieval~~
5. ~~Scaling ColPali to large collections~~
6. ~~RAG with VLMs~~
7. QA & Resources

# “Does it work for my data?”

Reply Guys
27.09.2024

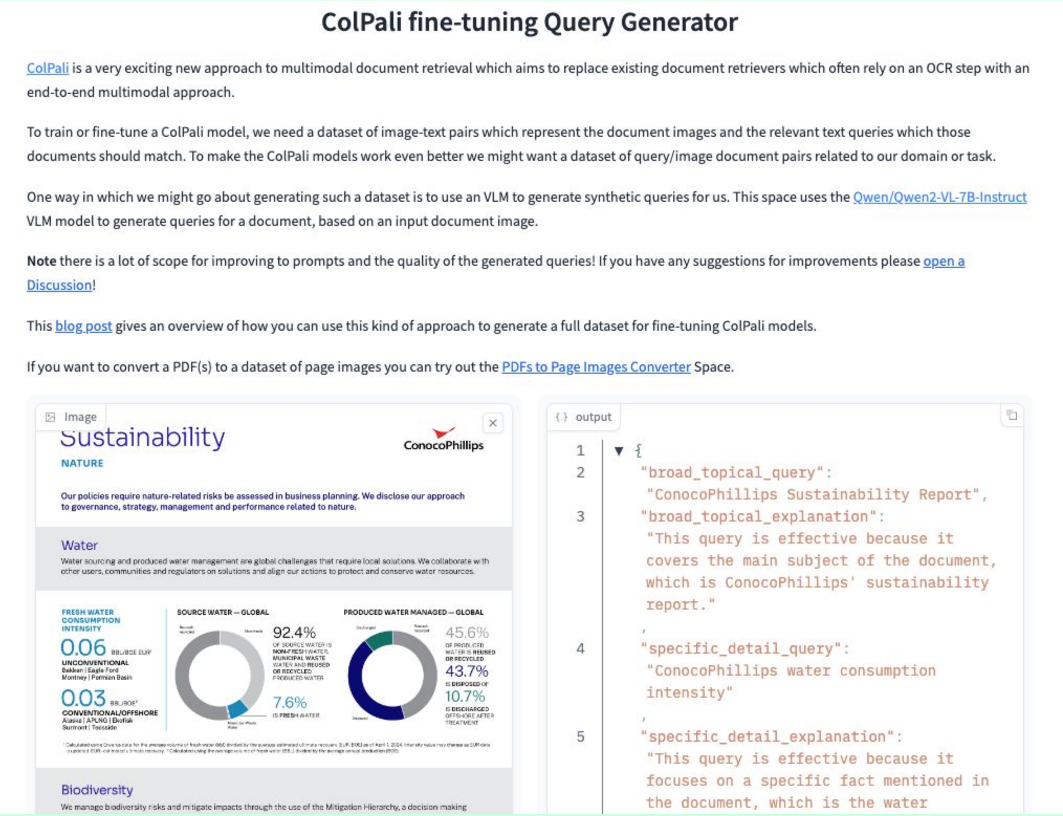
# Fine-tuning ColPali

[https://huggingface.co/spaces/davanstrien/ColPali-Query-Generator](https://huggingface.co/spaces/davanstrien/ColPali-Query-Generator)

ColPali **retrieval adapter** mostly trained on synthetic data created by previous generation VLMs using a prompting technique. Generate questions about a particular page.

New and improved VLM backbones gives:

- Better training data (generated on your domain and language)
- Better VLM backbone



## ColPali fine-tuning Query Generator

ColPali is a very exciting new approach to multimodal document retrieval which aims to replace existing document retrievers which often rely on an OCR step with an end-to-end multimodal approach.

To train or fine-tune a ColPali model, we need a dataset of image-text pairs which represent the document images and the relevant text queries which those documents should match. To make the ColPali models work even better we might want a dataset of query/image document pairs related to our domain or task.

One way in which we might go about generating such a dataset is to use an VLM to generate synthetic queries for us. This space uses the Qwen/Qwen2-VL-7B-Instruct VLM model to generate queries for a document, based on an input document image.

**Note** there is a lot of scope for improving to prompts and the quality of the generated queries! If you have any suggestions for improvements please open a Discussion!

This blog post gives an overview of how you can use this kind of approach to generate a full dataset for fine-tuning ColPali models.

If you want to convert a PDF(s) to a dataset of page images you can try out the PDFs to Page Images Converter Space.

Sustainability

ConocoPhillips

NATURE

Our policies require nature-related risks be assessed in business planning. We disclose our approach to governance, strategy, management and performance related to nature.

Water

Water sourcing and produced water management are global challenges that require local solutions. We collaborate with other users, communities and regulators on solutions and align our actions to protect and conserve water resources.

FRESH WATER CONSUMPTION INTENSITY
0.06 BBL/BOE EUR
UNCONVENTIONAL
Bakken | Eagle Ford
Montney | Permian Basin
0.03 BBL/BOE
CONVENTIONAL/OFFSHORE
Alaska | APLNG | Ekofisk
Surmont | Teesside
SOURCE WATER — GLOBAL
92.4%
OF SOURCE WATER IS NON-FRESH WATER, MUNICIPAL WASTE WATER AND REUSED OR RECYCLED PRODUCED WATER
7.6%
IS FRESH WATER
PRODUCED WATER MANAGED — GLOBAL
45.6%
OF PRODUCED WATER IS REUSED OR RECYCLED
43.7%
IS DISPOSED OF
10.7%
IS DISCHARGED OFFSHORE AFTER TREATMENT

Biodiversity

We manage biodiversity risks and mitigate impacts through the use of the Mitigation Hierarchy, a decision-making

```
{
  "broad_topical_query":
   "ConocoPhillips Sustainability Report",
  "broad_topical_explanation":
   "This query is effective because it
   covers the main subject of the document,
   which is ConocoPhillips' sustainability
   report."
  ,
  "specific_detail_query":
   "ConocoPhillips water consumption
   intensity"
  ,
  "specific_detail_explanation":
   "This query is effective because it
   focuses on a specific fact mentioned in
   the document, which is the water
```

# “Can’t you use GPT-4o for this?”

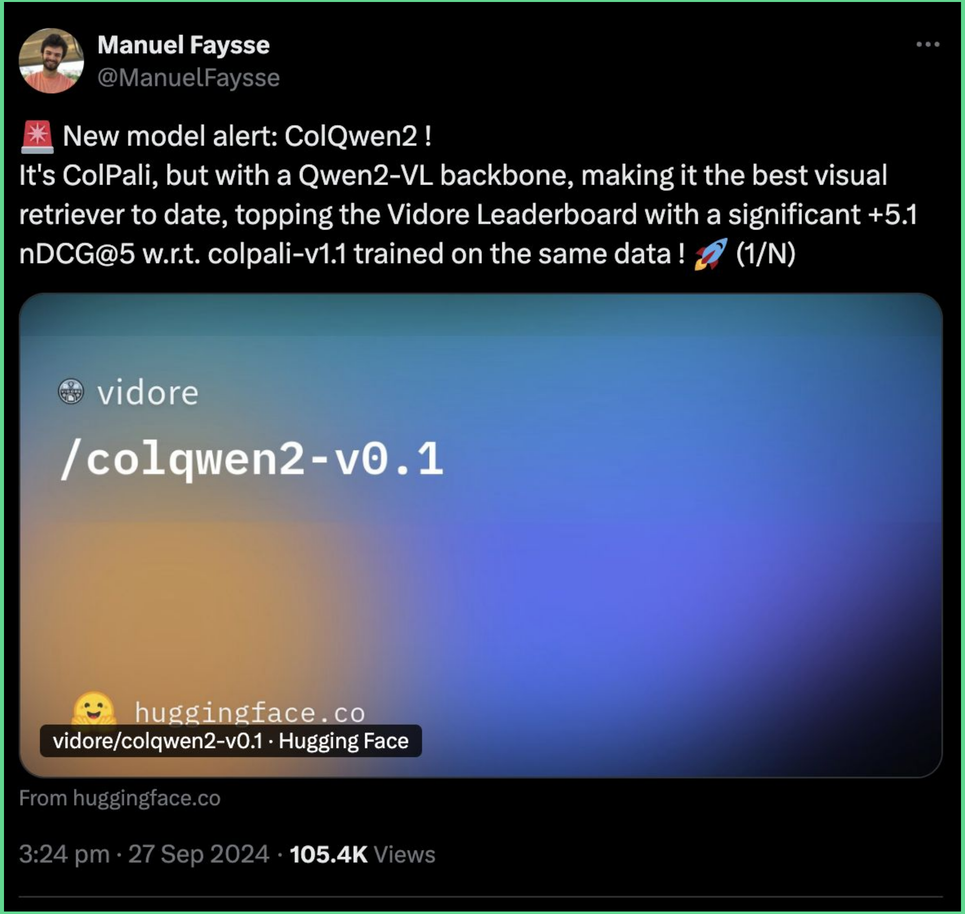
Reply Guys
27.09.2024

# ColPali is a promising direction

Will see new checkpoints based on new VLMs

Trained on more data

Embedding providers..



**Manuel Faysse**
@ManuelFaysse

🚨 New model alert: ColQwen2 !
It's ColPali, but with a Qwen2-VL backbone, making it the best visual retriever to date, topping the Vidore Leaderboard with a significant +5.1 nDCG@5 w.r.t. colpali-v1.1 trained on the same data ! 🚀 (1/N)

vidore/colqwen2-v0.1 · Hugging Face

From huggingface.co

3:24 pm · 27 Sep 2024 · **105.4K** Views

# QA

# Resources

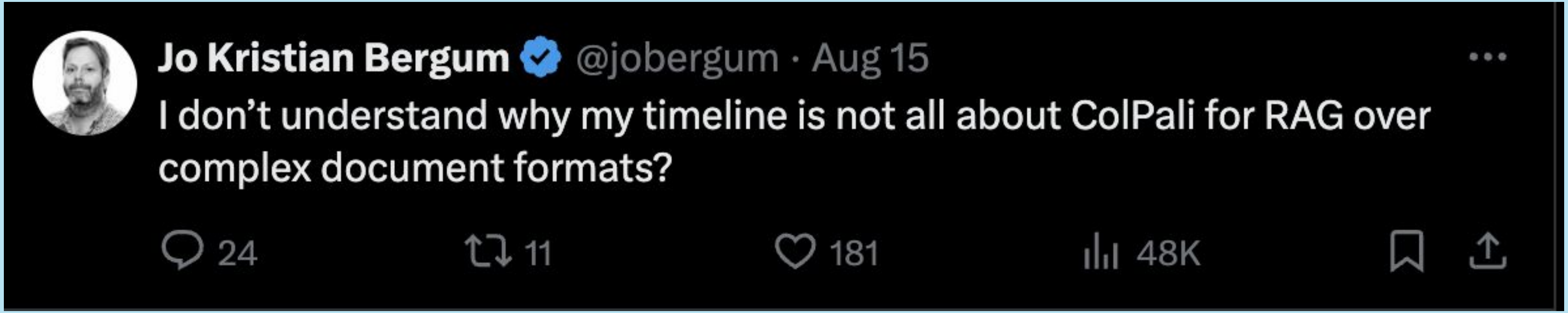
https://blog.vespa.ai/retrieval-with-vision-language-models-colpali/

https://blog.vespa.ai/the-rise-of-vision-driven-document-retrieval-for-rag/

https://blog.vespa.ai/scaling-colpali-to-billions/

https://huggingface.co/blog/manu/colpali

https://github.com/illuin-tech/colpali



**Jo Kristian Bergum** @jobergum · Aug 15

I don't understand why my timeline is not all about ColPali for RAG over complex document formats?

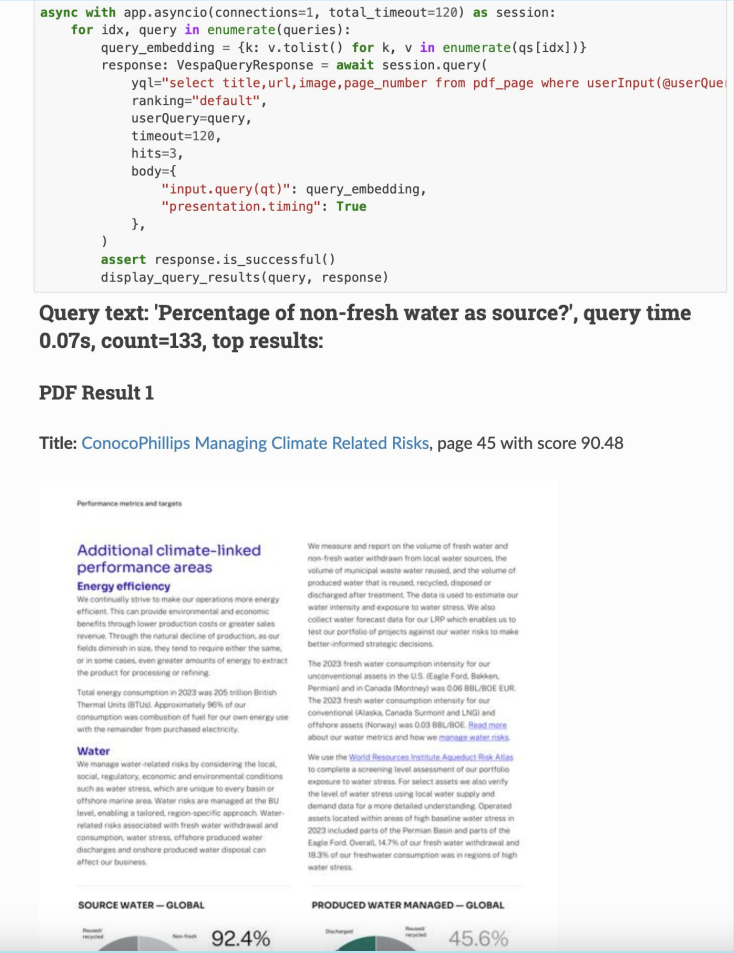
# Even more resources

[https://pyvespa.readthedocs.io/en/latest/examples/pdf-retrieval-with-ColQwen2-vlm_Vespa-cloud.html](https://pyvespa.readthedocs.io/en/latest/examples/pdf-retrieval-with-ColQwen2-vlm_Vespa-cloud.html)

[https://pyvespa.readthedocs.io/en/latest/examples/simplified-retrieval-with-colpali-vlm_Vespa-cloud.html](https://pyvespa.readthedocs.io/en/latest/examples/simplified-retrieval-with-colpali-vlm_Vespa-cloud.html)

[https://pyvespa.readthedocs.io/en/latest/examples/colpali-benchmark-vqa-vlm_Vespa-cloud.html](https://pyvespa.readthedocs.io/en/latest/examples/colpali-benchmark-vqa-vlm_Vespa-cloud.html)

[https://pyvespa.readthedocs.io/en/latest/examples/colpali-document-retrieval-vision-language-models-cloud.html](https://pyvespa.readthedocs.io/en/latest/examples/colpali-document-retrieval-vision-language-models-cloud.html)



```python
async with app.asyncio(connections=1, total_timeout=120) as session:
    for idx, query in enumerate(queries):
        query_embedding = {k: v.tolist() for k, v in enumerate(qs[idx])}
        response: VespaQueryResponse = await session.query(
            yql="select title,url,image,page_number from pdf_page where userInput(@userQue
            ranking="default",
            userQuery=query,
            timeout=120,
            hits=3,
            body={
                "input.query(qt)": query_embedding,
                "presentation.timing": True
            },
        )
        assert response.is_successful()
        display_query_results(query, response)
```

**Query text: 'Percentage of non-fresh water as source?', query time 0.07s, count=133, top results:**

**PDF Result 1**

**Title:** ConocoPhillips Managing Climate Related Risks, page 45 with score 90.48

Performance metrics and targets

Additional climate-linked performance areas

**Energy efficiency**

We continually strive to make our operations more energy efficient. This can provide environmental and economic benefits through lower production costs or greater sales revenue. Through the natural decline of production, as our fields diminish in size, they tend to require either the same, or in some cases, even greater amounts of energy to extract the product for processing or refining.

Total energy consumption in 2023 was 205 trillion British Thermal Units (BTUs). Approximately 96% of our consumption was combustion of fuel for our own energy use with the remainder from purchased electricity.

**Water**

We manage water-related risks by considering the local, social, regulatory, economic and environmental conditions such as water stress, which are unique to every basin or offshore marine area. Water risks are managed at the BU level, enabling a tailored, region-specific approach. Water-related risks associated with fresh water withdrawal and consumption, water stress, offshore produced water discharges and onshore produced water disposal can affect our business.

We measure and report on the volume of fresh water and non-fresh water withdrawn from local water sources, the volume of municipal waste water reused, and the volume of produced water that is reused, recycled, disposed or discharged after treatment. The data is used to estimate our water intensity and exposure to water stress. We also collect water forecast data for our LRP which enables us to test our portfolio of projects against our water risks to make better-informed strategic decisions.

The 2023 fresh water consumption intensity for our unconventional assets in the U.S. (Eagle Ford, Bakken, Permian) and in Canada (Montney) was 0.06 BBL/BOE EUR. The 2023 fresh water consumption intensity for our conventional (Alaska, Canada Surmont and LNG) and offshore assets (Norway) was 0.03 BBL/BOE. Read more about our water metrics and how we manage water risks.

We use the World Resources Institute Aqueduct Risk Atlas to complete a screening level assessment of our portfolio exposure to water stress. For select assets we also verify the level of water stress using local water supply and demand data for a more detailed understanding. Operated assets located within areas of high baseline water stress in 2023 included parts of the Permian Basin and parts of the Eagle Ford. Overall, 14.7% of our fresh water withdrawal and 18.3% of our freshwater consumption was in regions of high water stress.

**SOURCE WATER — GLOBAL**

**PRODUCED WATER MANAGED — GLOBAL**

Vespa

Follow us!

Vespa.ai
@vespaengine
linkedin.com/company/vespa-ai/