GOSIM AI Paris 2025

RAGFlow—Leading the Open-Source Revolution in Enterprise-Grade RAG

Yingfeng@InfiniFlow.ai

① Introduction

② Chunking

③ Retrieval

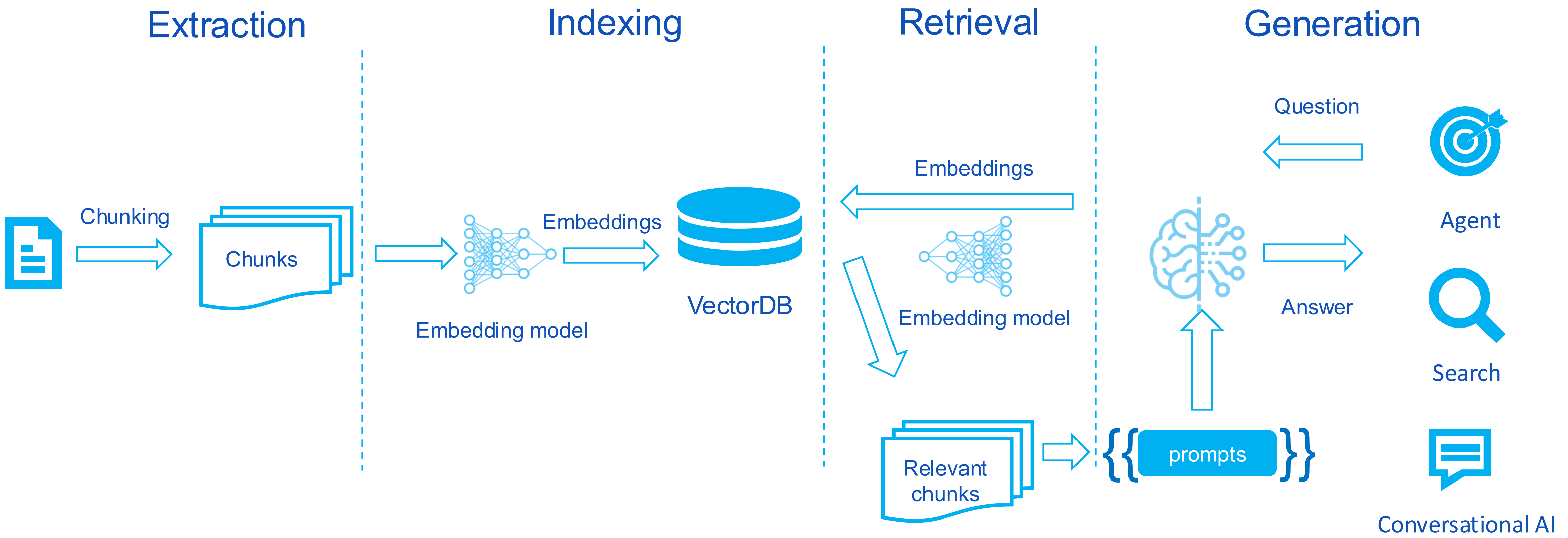
④ Agentic and future

⑤ Summary

Introduction

Paradigm of RAG

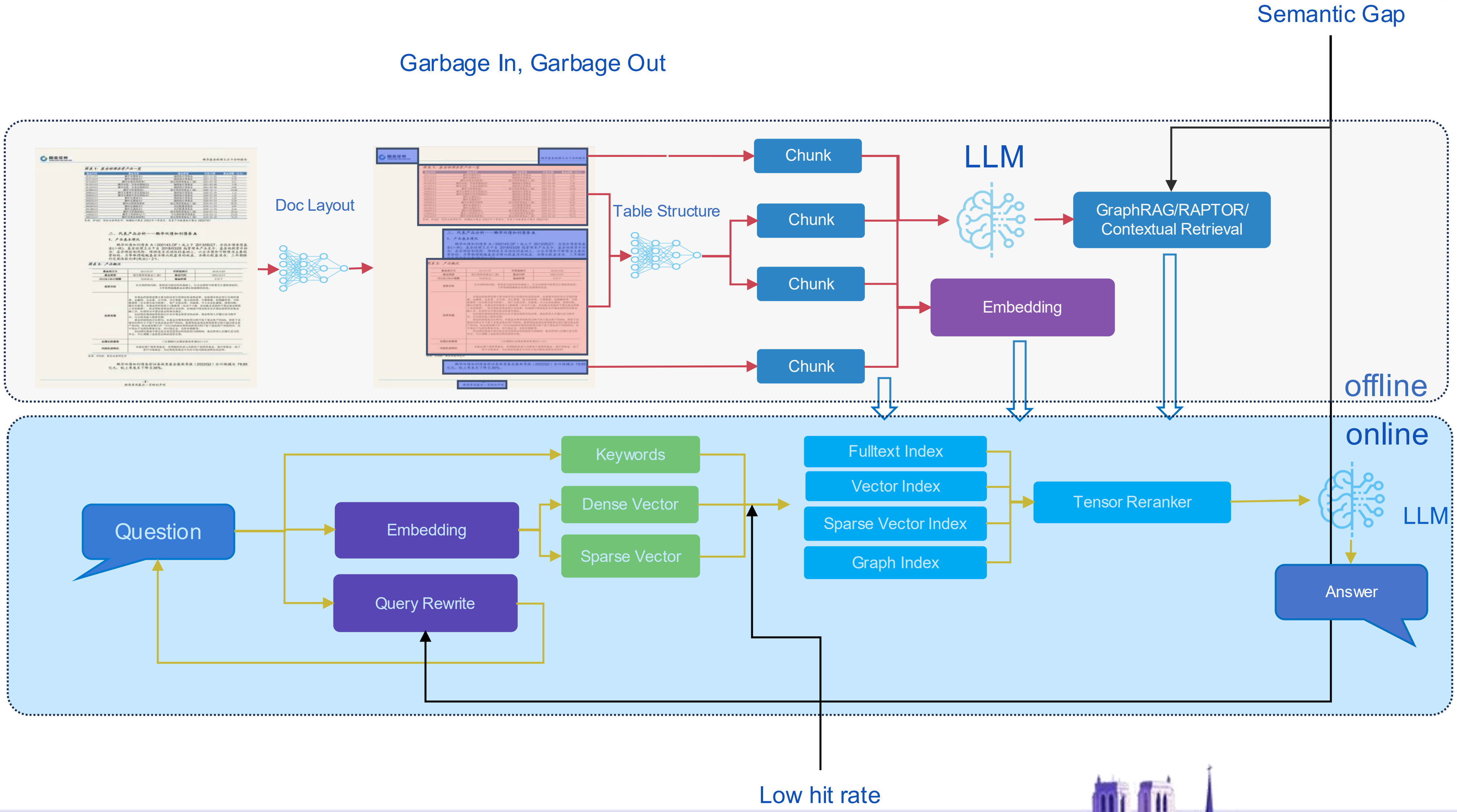
GOSIM



Challenges of RAG

- ❌ Garbage in, garbage out
- ❌ Low hit rate
- ❌ Semantic gap between question and answer

Architecture of RAGFlow



Semantic Gap

Garbage In, Garbage Out

Doc Layout

Table Structure

Chunk

Chunk

Chunk

Chunk

LLM

GraphRAG/RAPTOR/
Contextual Retrieval

Embedding

offline

online

Question

Embedding

Query Rewrite

Keywords

Dense Vector

Sparse Vector

Fulltext Index

Vector Index

Sparse Vector Index

Graph Index

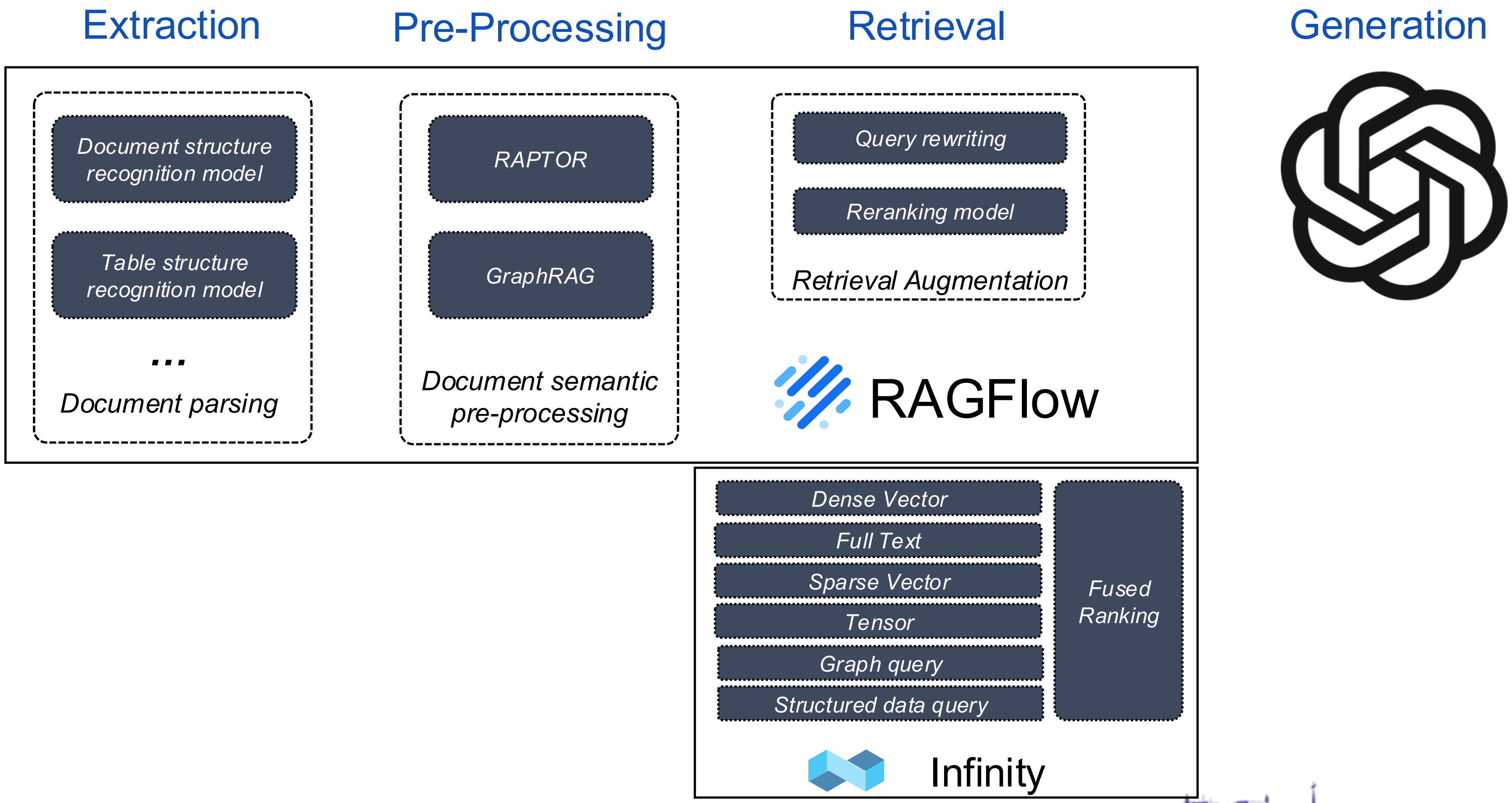
Tensor Reranker

LLM

Answer

Low hit rate

RAG ≠ LLMOps



Extraction

Document structure recognition model

Table structure recognition model

…

Document parsing

Pre-Processing

RAPTOR

GraphRAG

Document semantic pre-processing

Retrieval

Query rewriting

Reranking model

Retrieval Augmentation

RAGFlow

Dense Vector

Full Text

Sparse Vector

Tensor

Graph query

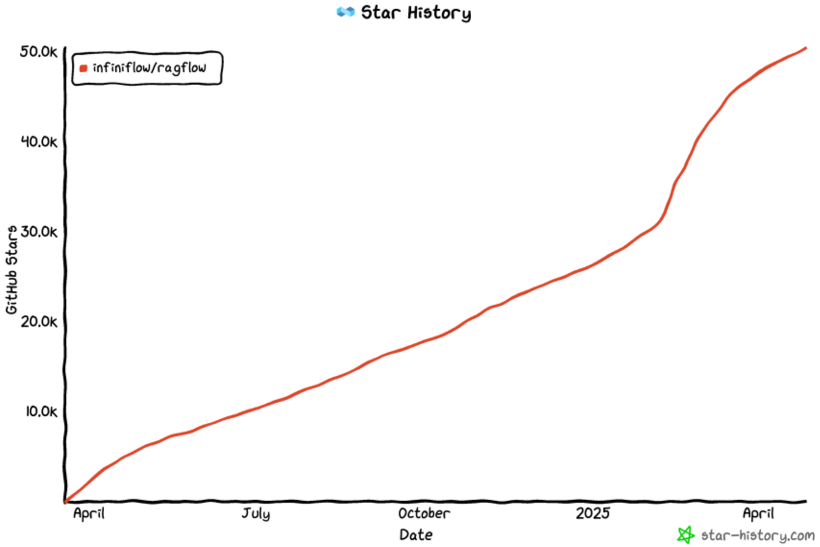
Structured data query

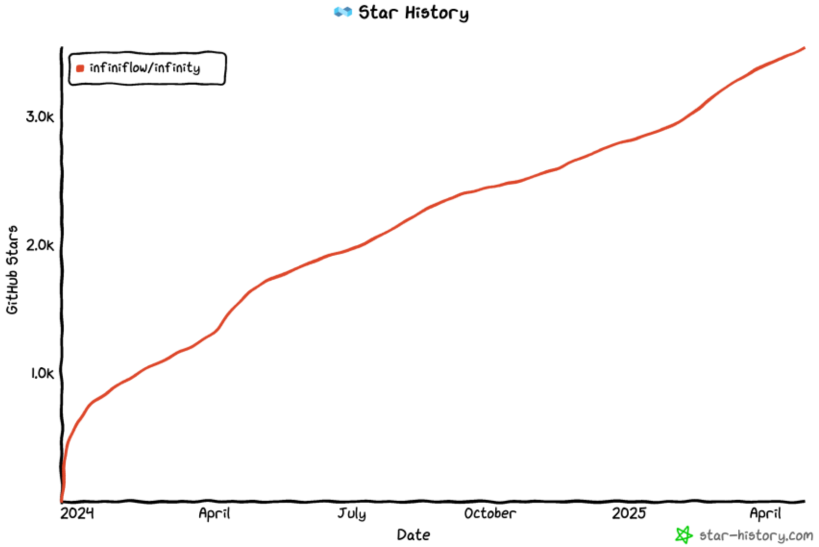
Fused Ranking

Infinity

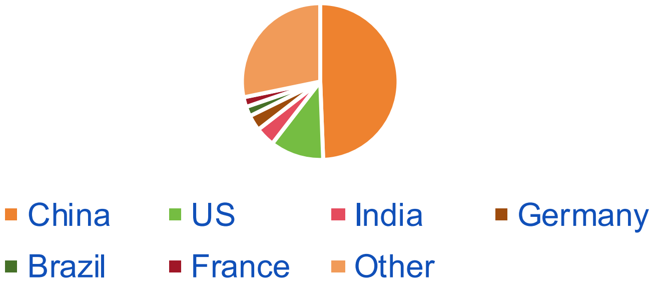
Generation

InfiniFlow = Infinity + RAGFlow

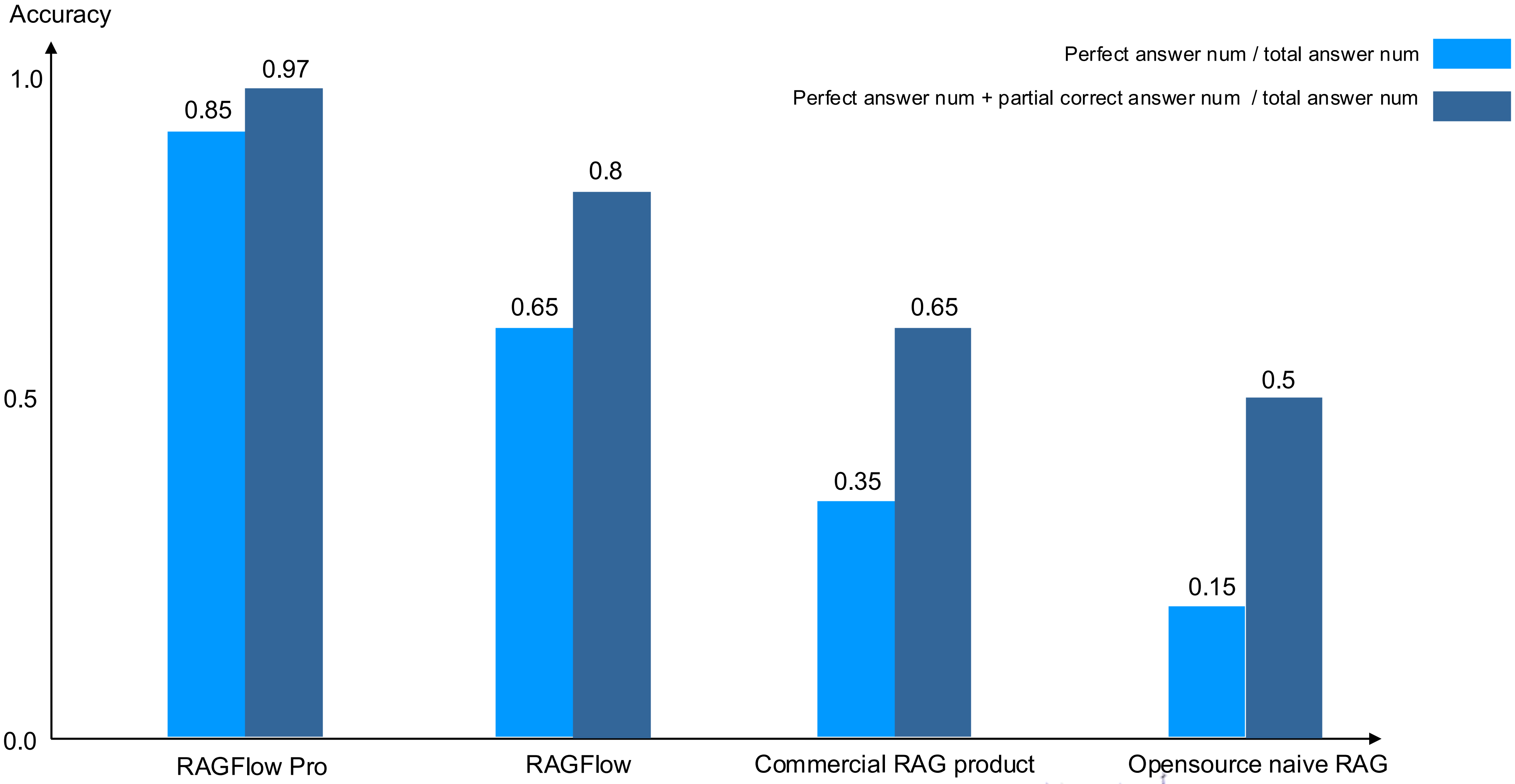


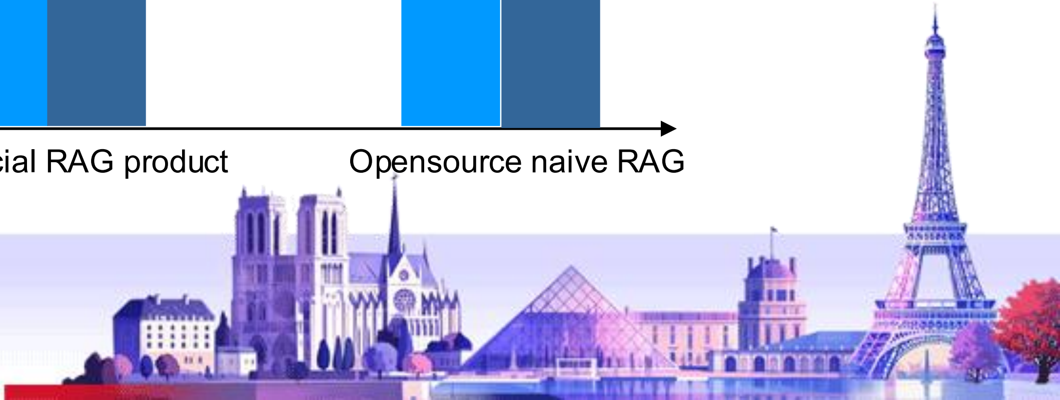


Top Star Geo-Locations



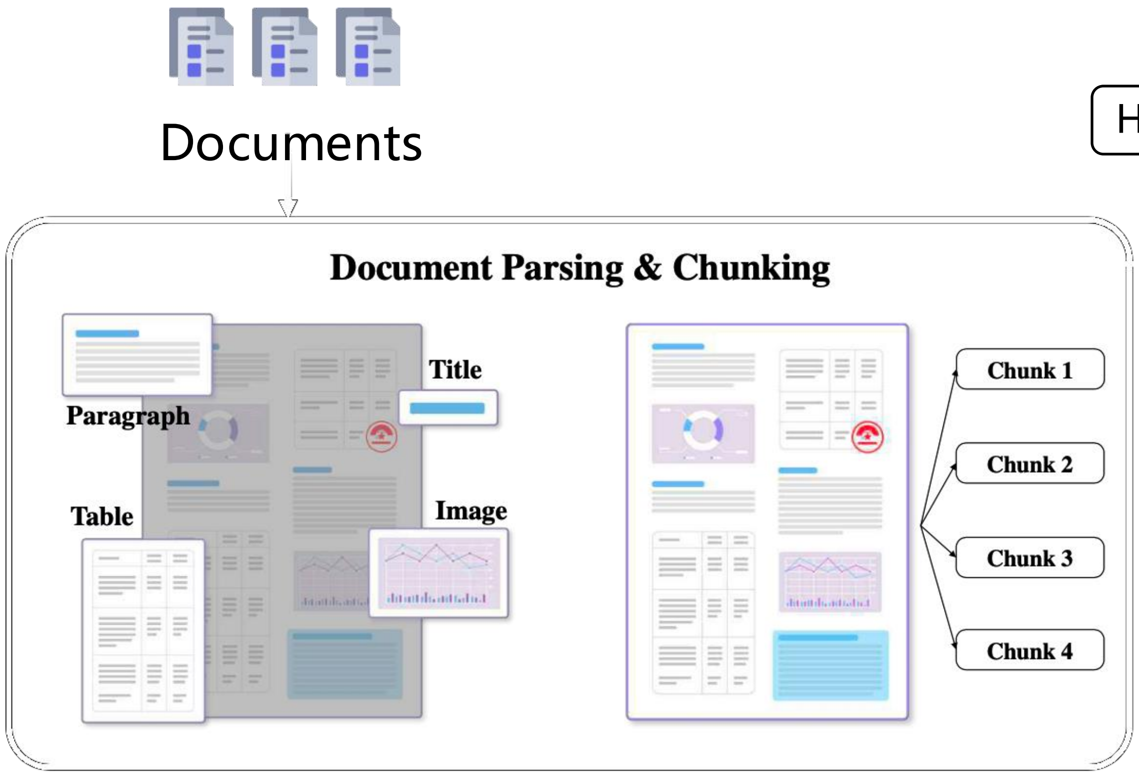
RAGFlow vs Others

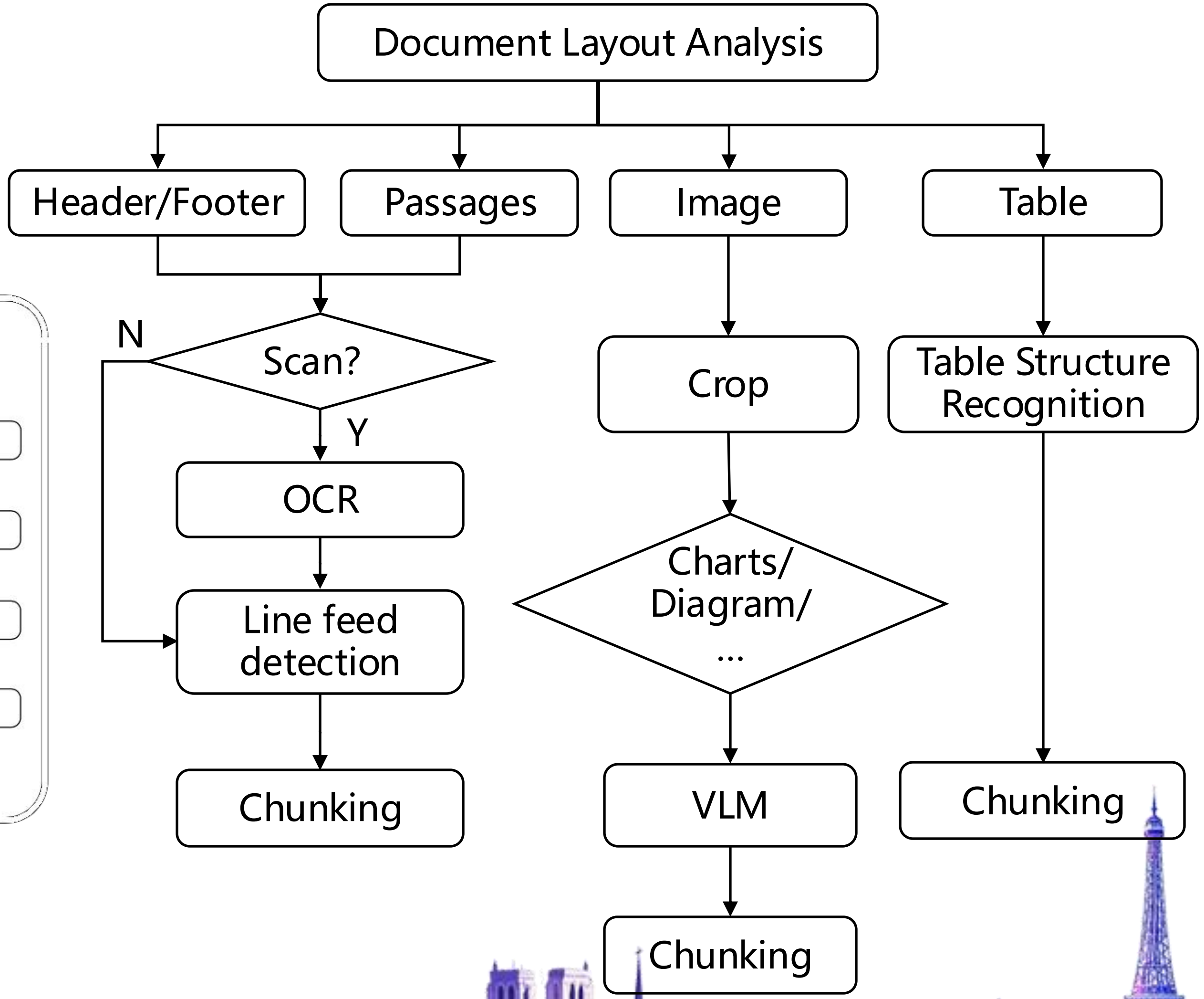




Chunking

Deep Document Understanding





Document Layout Analysis

Header/Footer

Passages

Image

Table

Scan?

N

Y

OCR

Line feed detection

Chunking

Crop

Charts/ Diagram/ …

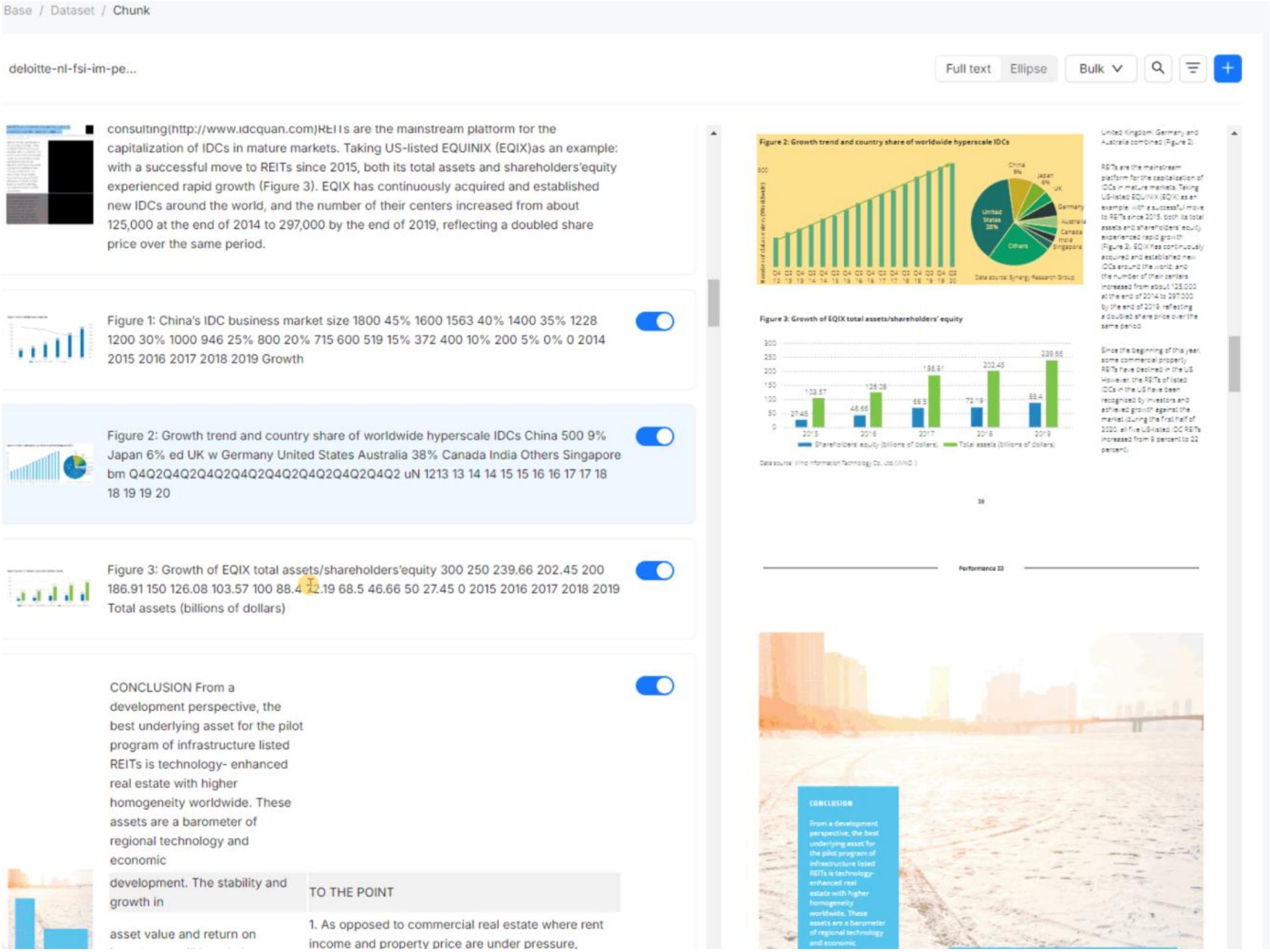
VLM

Chunking

Table Structure Recognition

Chunking

Visual Chunking Results


Base / Dataset / Chunk
deloitte-nl-fsi-im-pe...
Full text
Ellipse
Bulk
+
Figure 2: Growth trend and country share of worldwide hyperscale IDCs
China 9%
Japan 6%
UK
Germany
United States 38%
Australia
Canada
India
Singapore
Others
Data source: Synergy Research Group
Figure 3: Growth of EQIX total assets/shareholders' equity
Shareholders' equity (billions of dollars)
Total assets (billions of dollars)
CONCLUSION

consulting(http://www.idcquan.com)REITs are the mainstream platform for the capitalization of IDCs in mature markets. Taking US-listed EQUINIX (EQIX)as an example: with a successful move to REITs since 2015, both its total assets and shareholders'equity experienced rapid growth (Figure 3). EQIX has continuously acquired and established new IDCs around the world, and the number of their centers increased from about 125,000 at the end of 2014 to 297,000 by the end of 2019, reflecting a doubled share price over the same period.

Figure 1: China's IDC business market size 1800 45% 1600 1563 40% 1400 35% 1228 1200 30% 1000 946 25% 800 20% 715 600 519 15% 372 400 10% 200 5% 0% 0 2014 2015 2016 2017 2018 2019 Growth

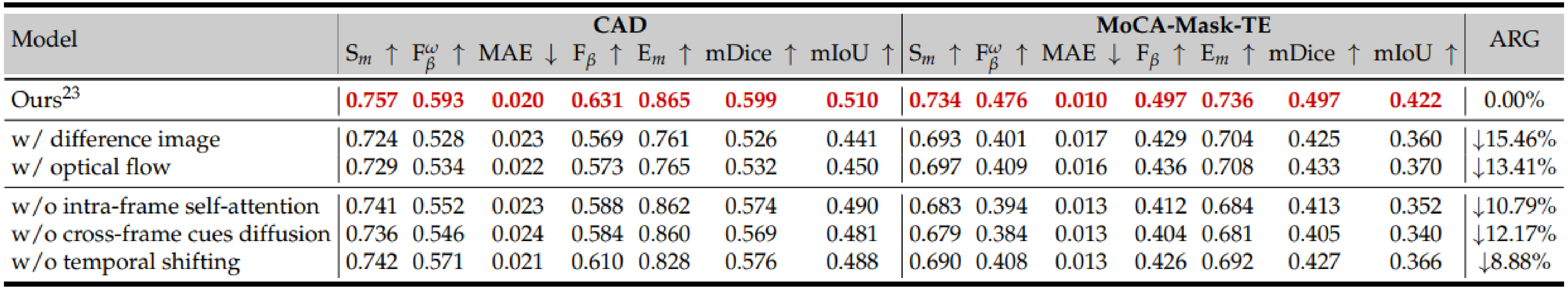
Figure 2: Growth trend and country share of worldwide hyperscale IDCs China 500 9% Japan 6% ed UK w Germany United States Australia 38% Canada India Others Singapore bm Q4Q2Q4Q2Q4Q2Q4Q2Q4Q2Q4Q2Q4Q2Q4Q2 uN 1213 13 14 14 15 15 16 16 17 17 18 18 19 19 20

Figure 3: Growth of EQIX total assets/shareholders'equity 300 250 239.66 202.45 200 186.91 150 126.08 103.57 100 88.4 72.19 68.5 46.66 50 27.45 0 2015 2016 2017 2018 2019 Total assets (billions of dollars)

CONCLUSION From a development perspective, the best underlying asset for the pilot program of infrastructure listed REITs is technology- enhanced real estate with higher homogeneity worldwide. These assets are a barometer of regional technology and economic

development. The stability and growth in	TO THE POINT
asset value and return on	1. As opposed to commercial real estate where rent income and property price are under pressure,

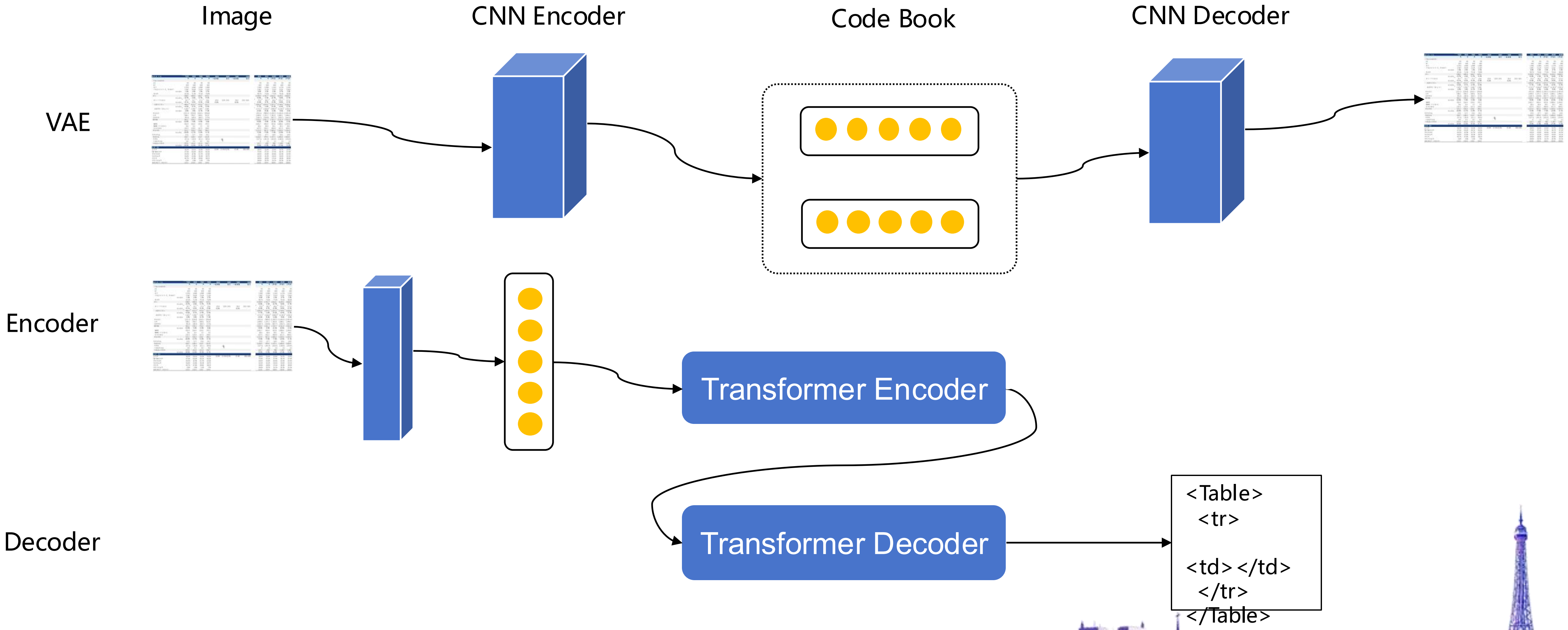
Table Structure Recognition



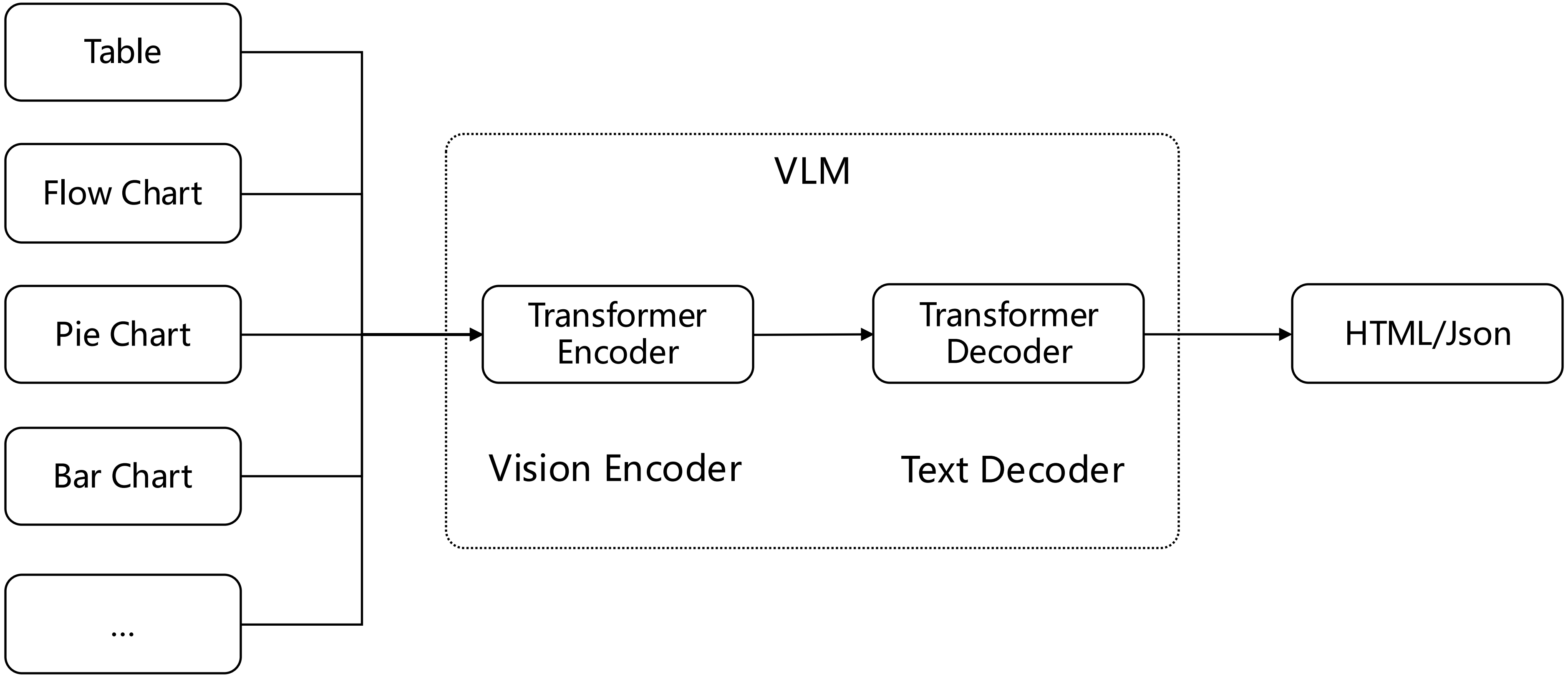
Model	CAD							MoCA-Mask-TE							ARG
	S_m ↑	F^{ω}_{β} ↑	MAE ↓	F_β ↑	E_m ↑	mDice ↑	mIoU ↑	S_m ↑	F^{ω}_{β} ↑	MAE ↓	F_β ↑	E_m ↑	mDice ↑	mIoU ↑	
Ours[23]	**0.757**	**0.593**	**0.020**	**0.631**	**0.865**	**0.599**	**0.510**	**0.734**	**0.476**	**0.010**	**0.497**	**0.736**	**0.497**	**0.422**	0.00%
w/ difference image	0.724	0.528	0.023	0.569	0.761	0.526	0.441	0.693	0.401	0.017	0.429	0.704	0.425	0.360	↓15.46%
w/ optical flow	0.729	0.534	0.022	0.573	0.765	0.532	0.450	0.697	0.409	0.016	0.436	0.708	0.433	0.370	↓13.41%
w/o intra-frame self-attention	0.741	0.552	0.023	0.588	0.862	0.574	0.490	0.683	0.394	0.013	0.412	0.684	0.413	0.352	↓10.79%
w/o cross-frame cues diffusion	0.736	0.546	0.024	0.584	0.860	0.569	0.481	0.679	0.384	0.013	0.404	0.681	0.405	0.340	↓12.17%
w/o temporal shifting	0.742	0.571	0.021	0.610	0.828	0.576	0.488	0.690	0.408	0.013	0.426	0.692	0.427	0.366	↓8.88%

- Multi table header
- Border/Bordless
- Cell merge
- Cross page

DeepDoc—CNN or GenAI?



DeepDoc—CNN or GenAI?



Table

Flow Chart

Pie Chart

Bar Chart

...

VLM

Transformer Encoder

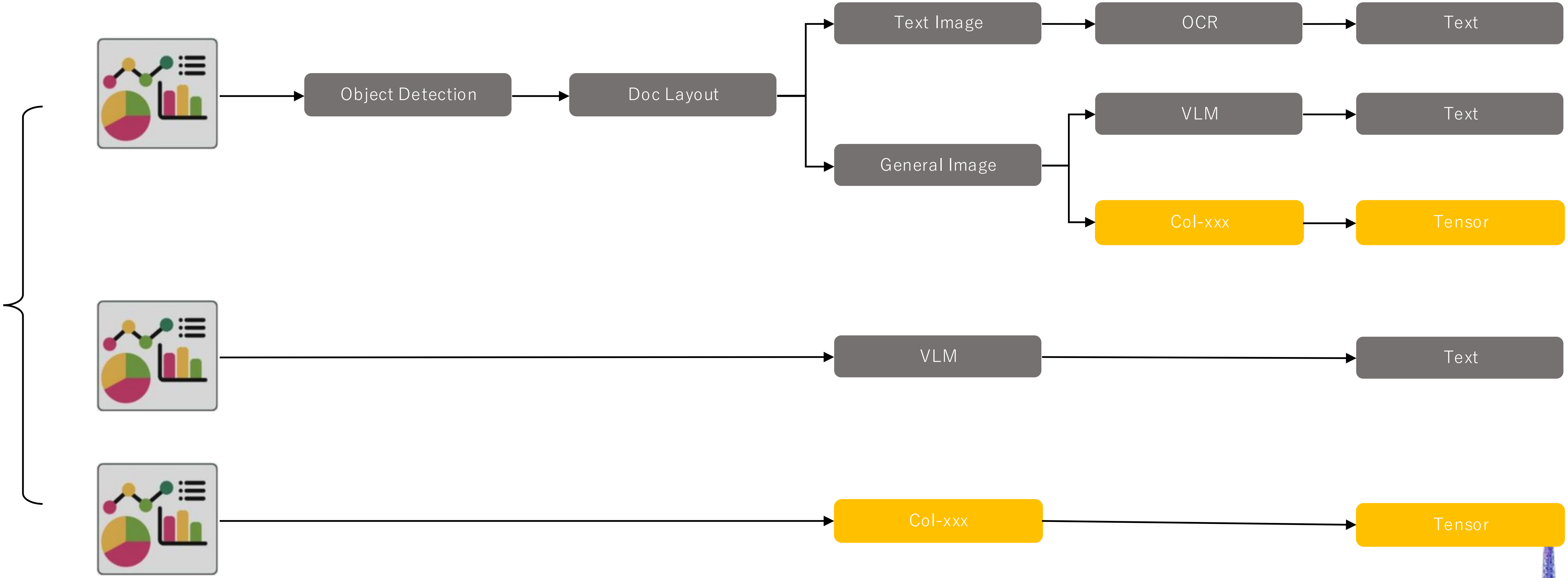
Transformer Decoder

Vision Encoder

Text Decoder

HTML/Json

Multi-Modal RAG



Object Detection

Doc Layout

Text Image

OCR

Text

General Image

VLM

Text

Col-xxx

Tensor

VLM

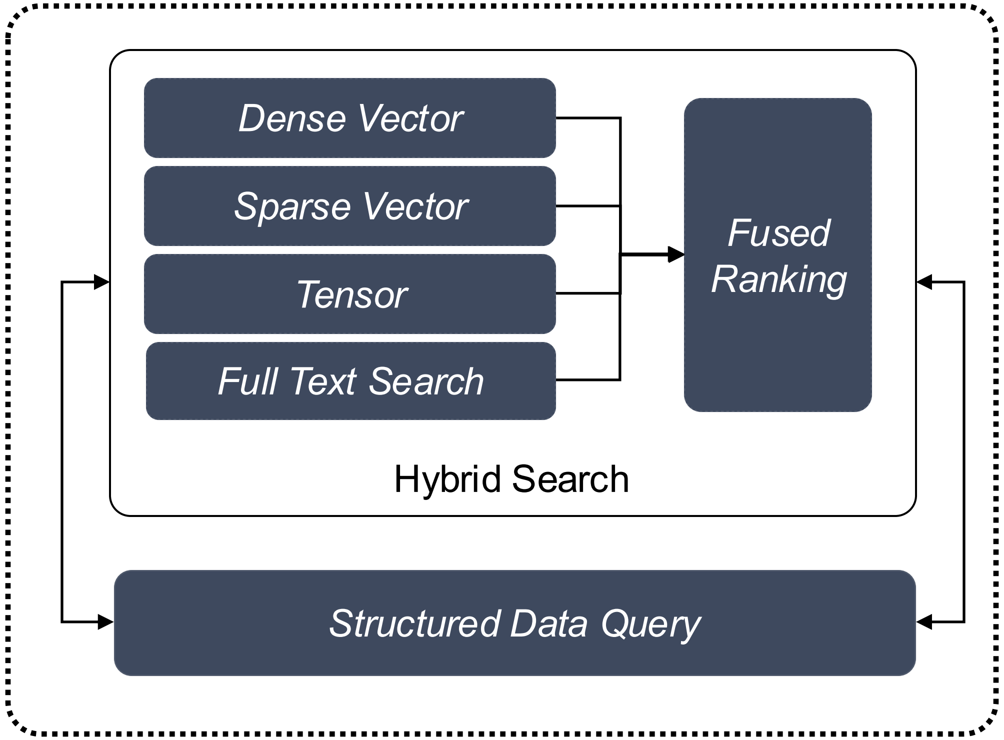
Text

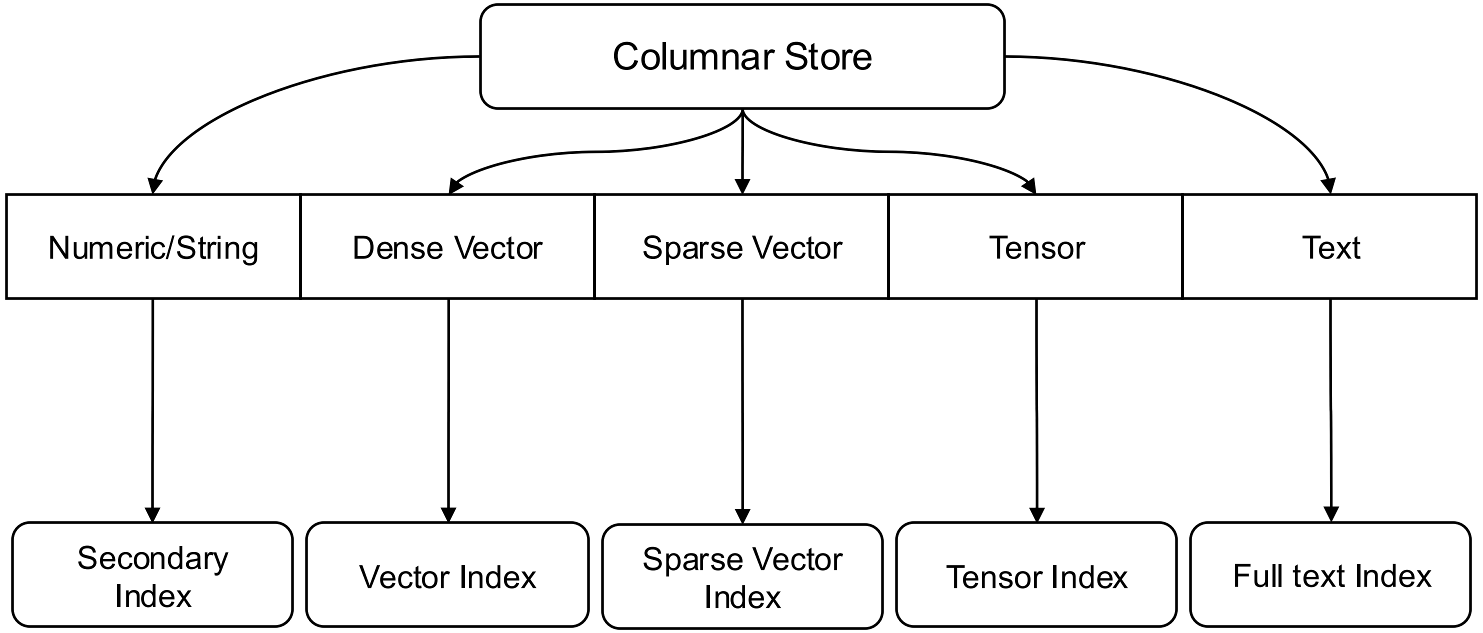
Col-xxx

Tensor

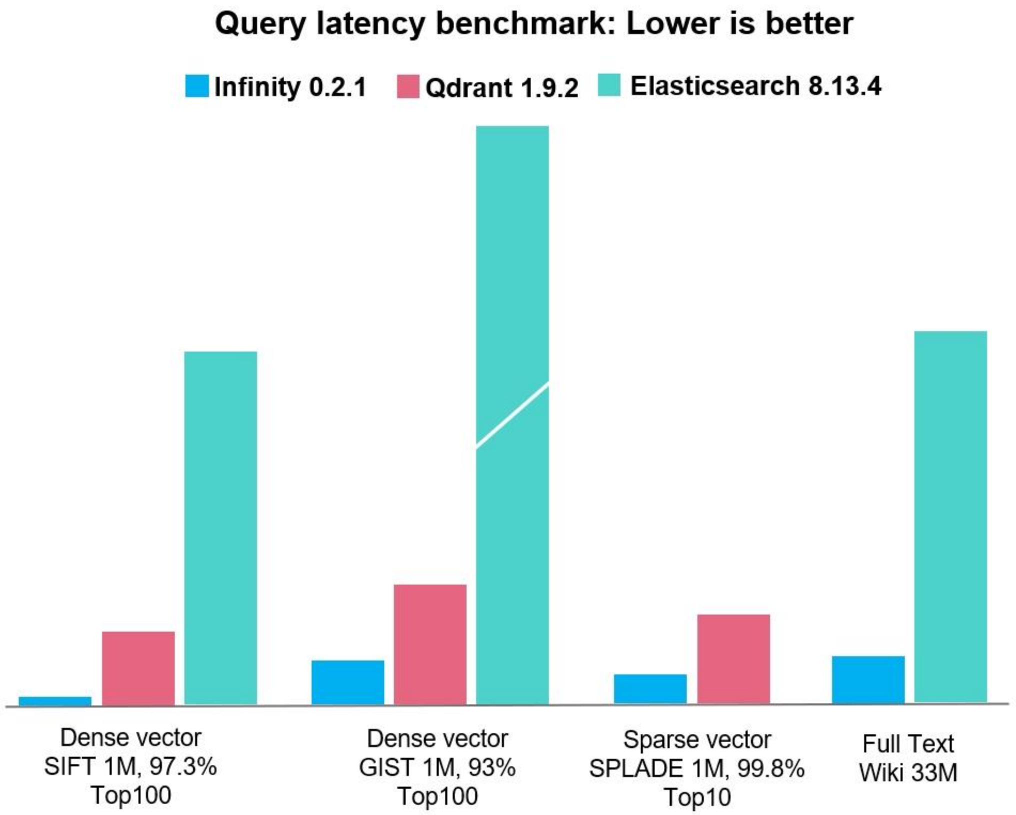
Retrieval

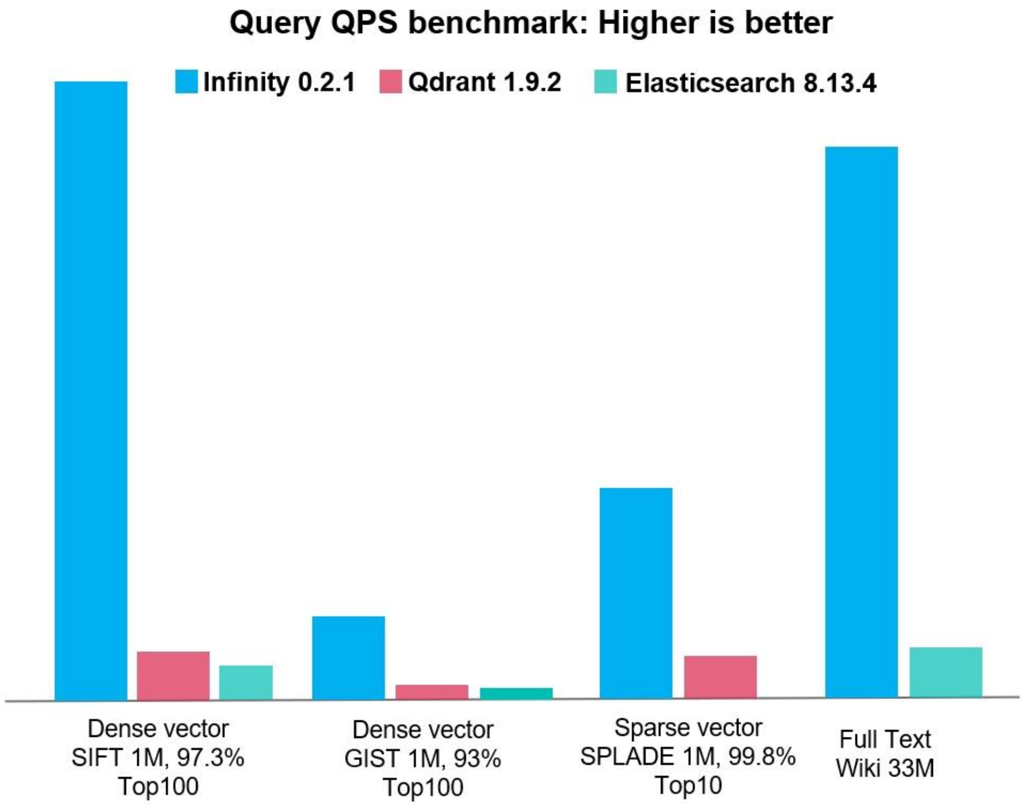
Infinity—Indexing Database



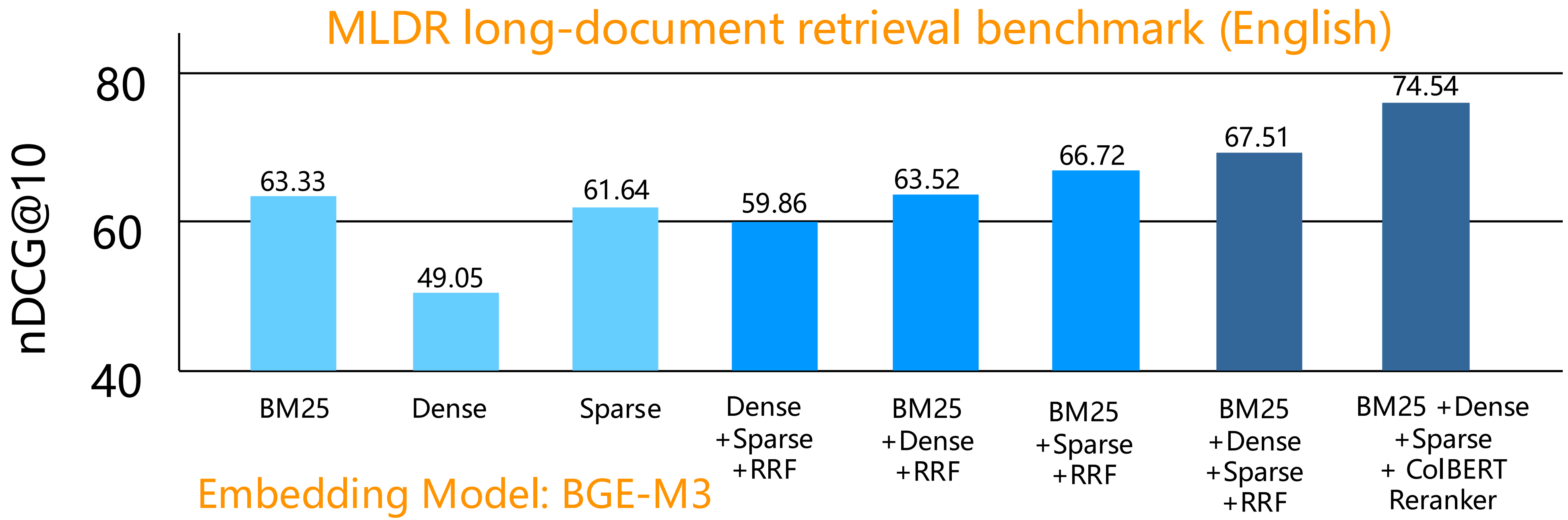


Benchmark


Query latency benchmark: Lower is better
Infinity 0.2.1
Qdrant 1.9.2
Elasticsearch 8.13.4
Dense vector SIFT 1M, 97.3% Top100
Dense vector GIST 1M, 93% Top100
Sparse vector SPLADE 1M, 99.8% Top10
Full Text Wiki 33M


Query QPS benchmark: Higher is better
Infinity 0.2.1
Qdrant 1.9.2
Elasticsearch 8.13.4
Dense vector SIFT 1M, 97.3% Top100
Dense vector GIST 1M, 93% Top100
Sparse vector SPLADE 1M, 99.8% Top10
Full Text Wiki 33M

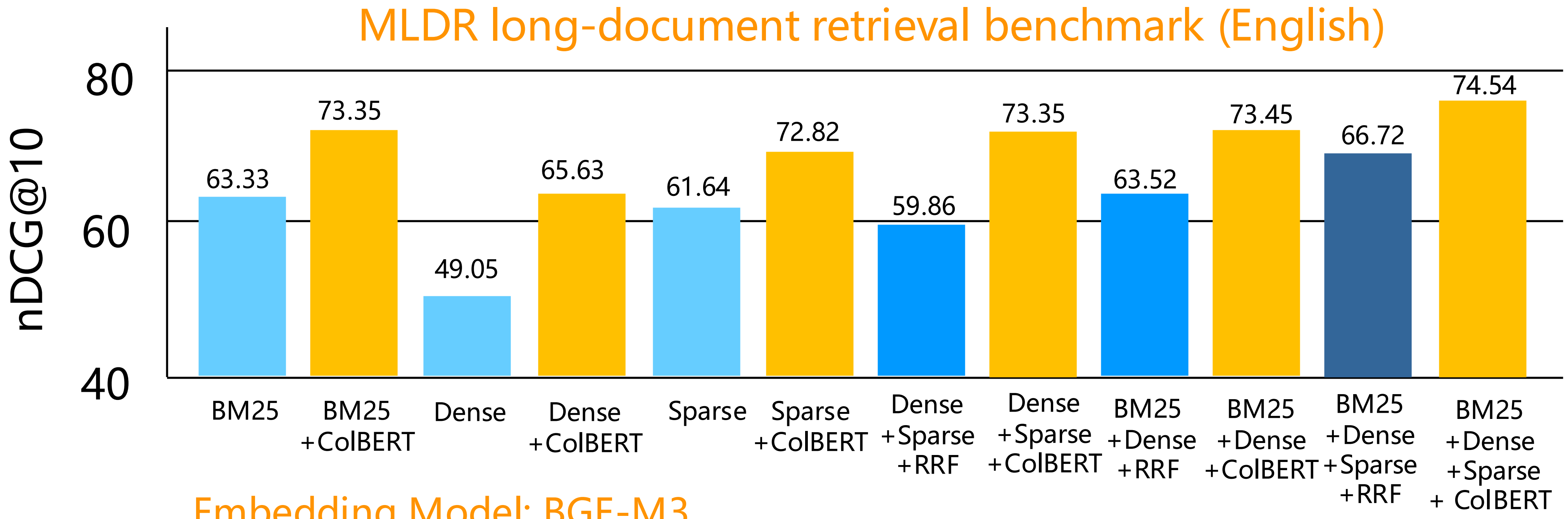
Hybrid Search



Embedding Model: BGE-M3

IBM Blended RAG
https://arxiv.org/abs/2404.07220

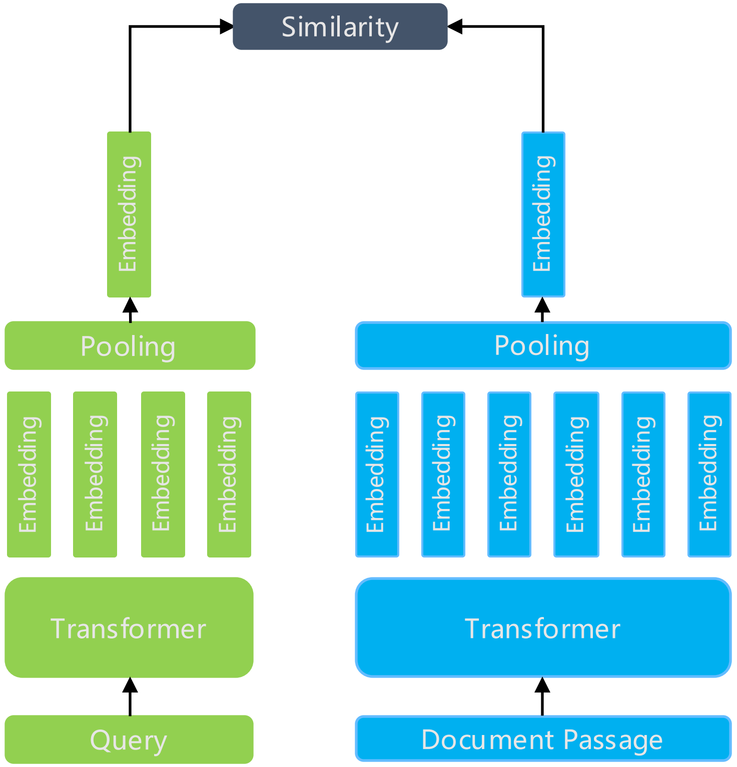
Benefits of ColBERT



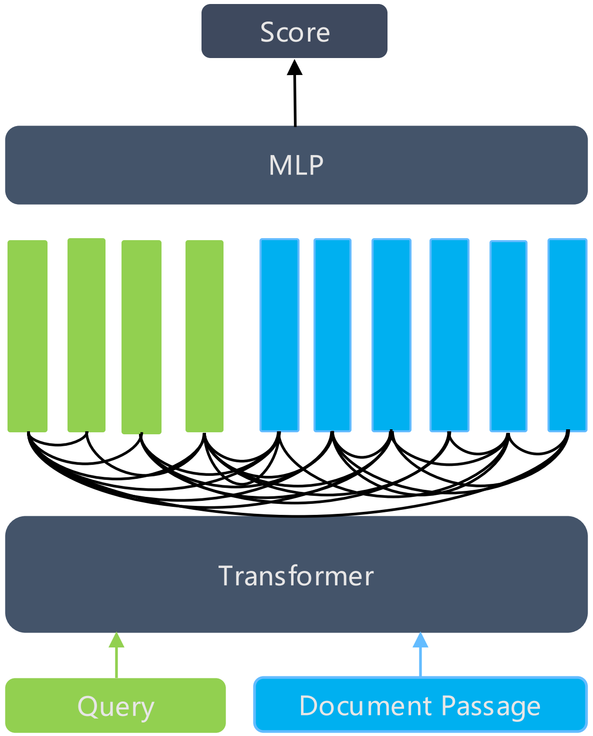
Embedding Model: BGE-M3

Ranking Model

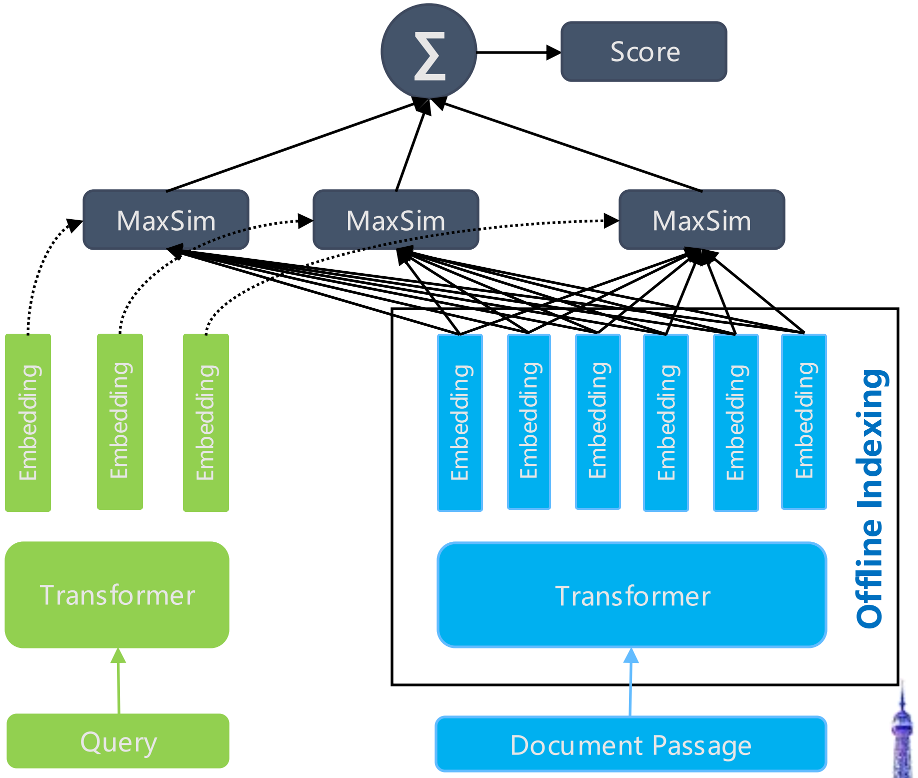
Dual Encoder


Similarity
Embedding
Embedding
Pooling
Pooling
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Transformer
Transformer
Query
Document Passage

Cross Encoder

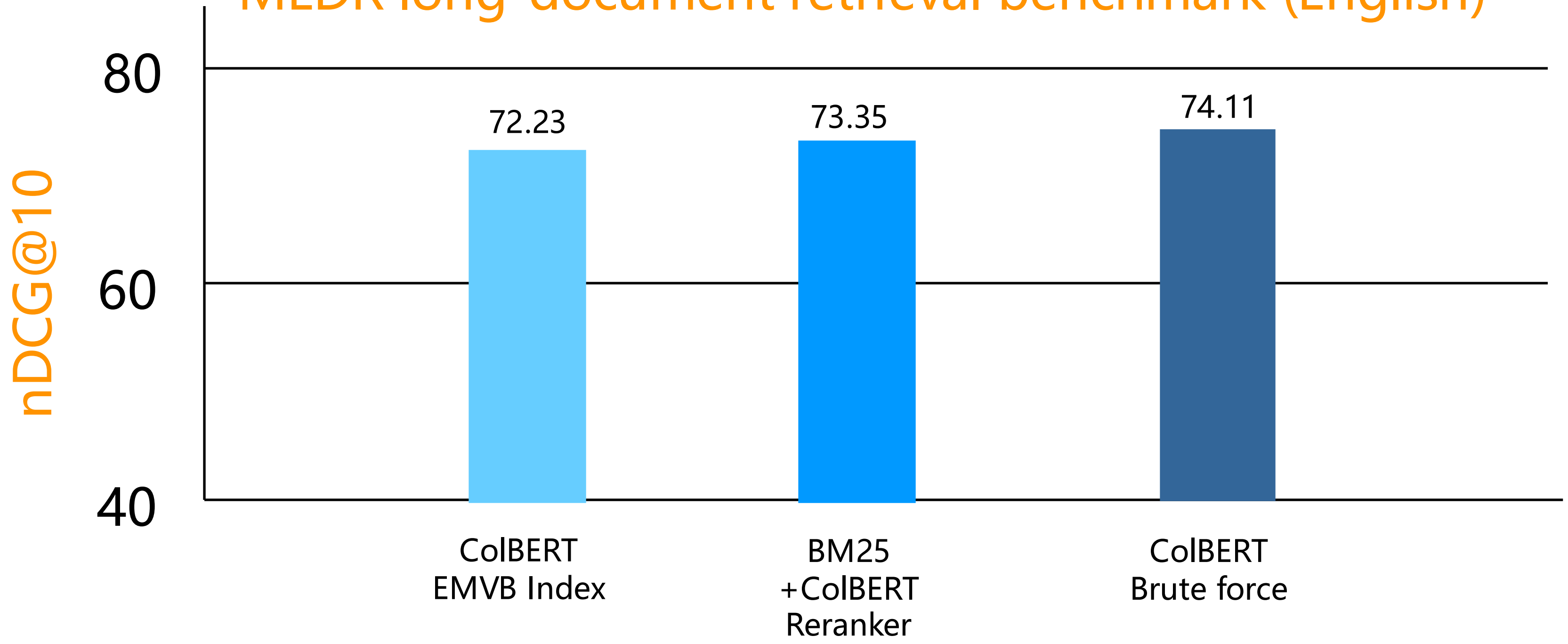

Score
MLP
Transformer
Query
Document Passage

Late Interaction Encoder
ColBERT


Σ
Score
MaxSim
MaxSim
MaxSim
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Embedding
Offline Indexing
Transformer
Transformer
Query
Document Passage

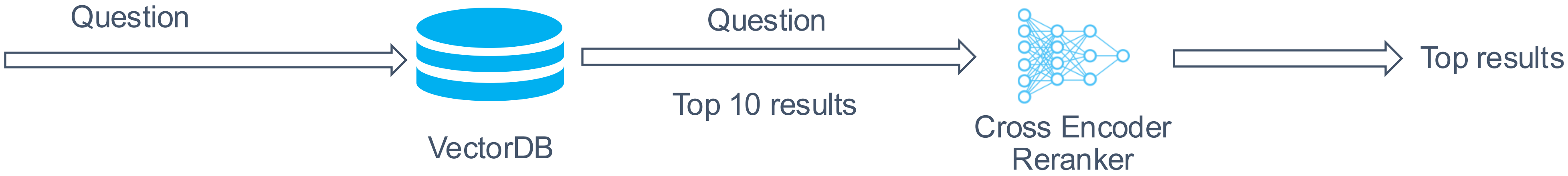
ColBERT Ranker vs Reranker

MLDR long-document retrieval benchmark (English)



Embedding Model: BGE-M3

Benefits of Col-xxx Rerankers



Question

VectorDB

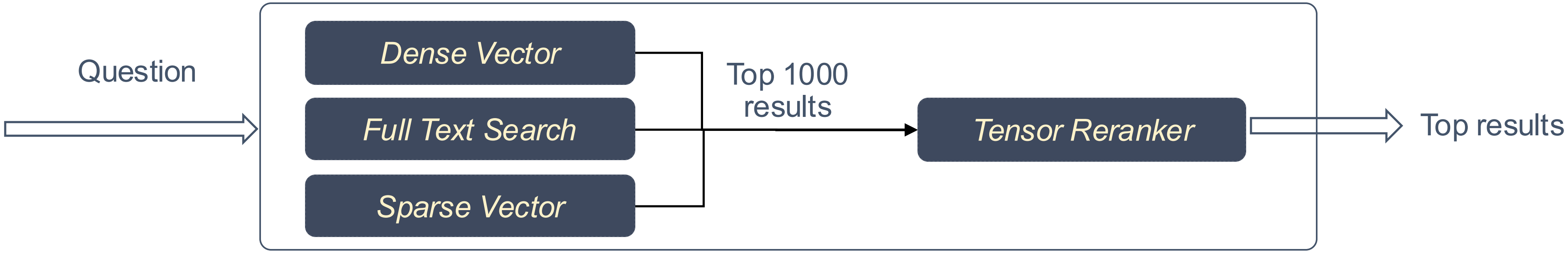
Question

Top 10 results

Cross Encoder Reranker

Top results

VS



Question

Dense Vector

Full Text Search

Sparse Vector

Top 1000 results

Tensor Reranker

Top results

Col-xxx in Multi-Modal-RAG

Divides a PDF page into 32 × 32 = 1024 patches

Each patch is represented as a 128-dim vector

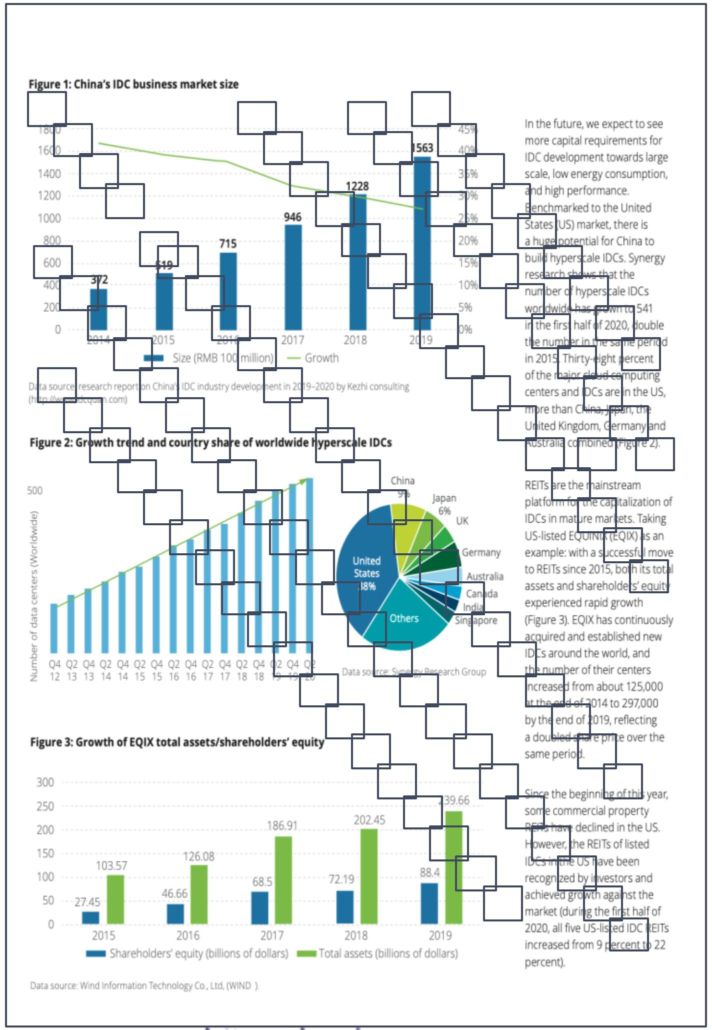


Figure 1: China's IDC business market size

Data source: research report on China's IDC industry development in 2019–2020 by Kezhi consulting (http://www.idcquan.com)

Figure 2: Growth trend and country share of worldwide hyperscale IDCs

Data source: Synergy Research Group

Figure 3: Growth of EQIX total assets/shareholders' equity

27.45
103.57
46.66
126.08
68.5
186.91
72.19
202.45
88.4
239.66
2015
2016
2017
2018
2019
Shareholders' equity (billions of dollars)
Total assets (billions of dollars)

Data source: Wind Information Technology Co., Ltd, (WIND).

In the future, we expect to see more capital requirements for IDC development towards large scale, low energy consumption, and high performance. Benchmarked to the United States (US) market, there is a huge potential for China to build hyperscale IDCs. Synergy research shows that the number of hyperscale IDCs worldwide has grown to 541 in the first half of 2020, double the number in the same period in 2015. Thirty-eight percent of the major cloud computing centers and IDCs are in the US, more than China, Japan, the United Kingdom, Germany and Australia combined (Figure 2).

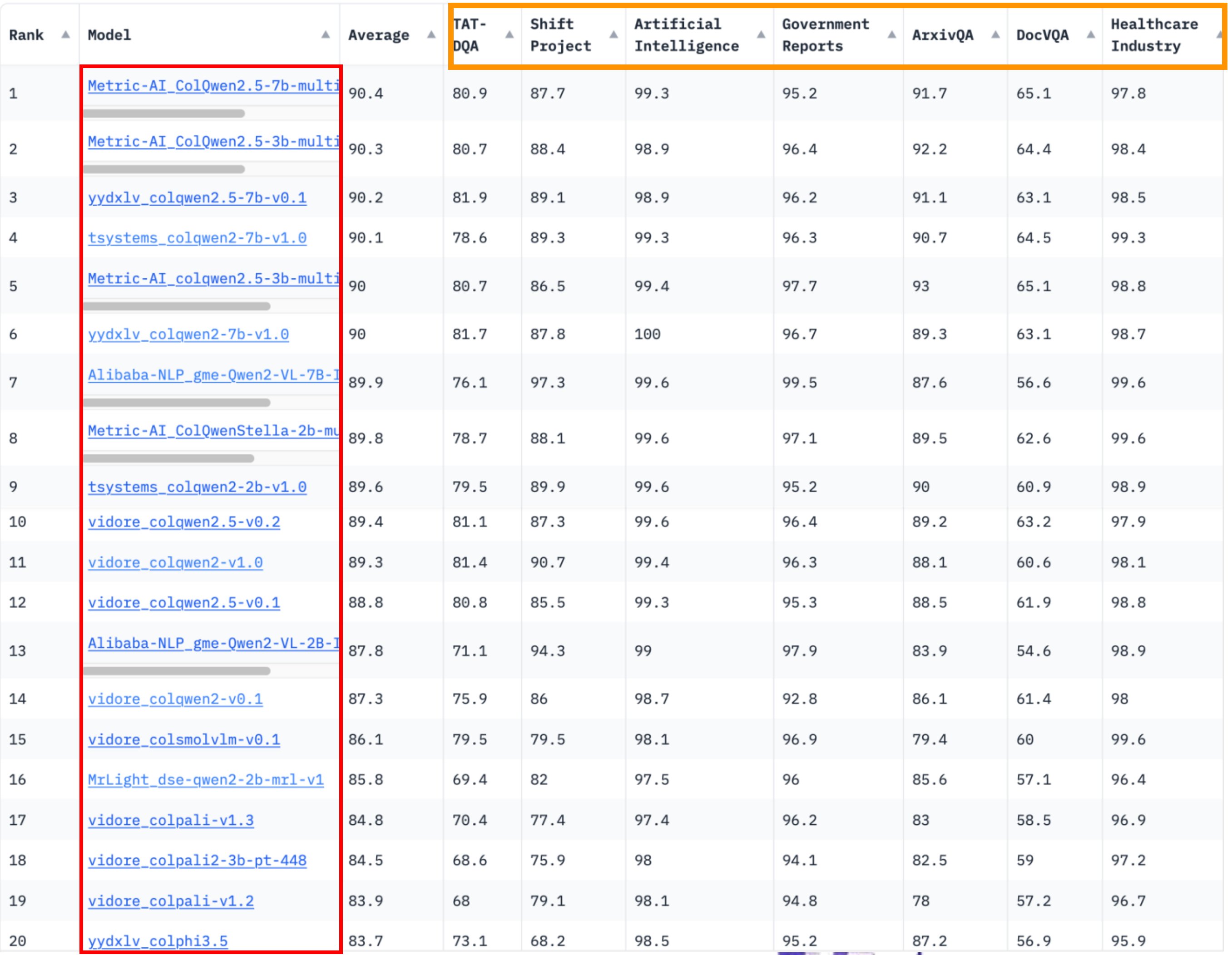
REITs are the mainstream platform for the capitalization of IDCs in mature markets. Taking US-listed EQUINIX (EQIX) as an example: with a successful move to REITs since 2015, both its total assets and shareholders' equity experienced rapid growth (Figure 3). EQIX has continuously acquired and established new IDCs around the world, and the number of their centers increased from about 125,000 at the end of 2014 to 297,000 by the end of 2019, reflecting a doubled share price over the same period.

Since the beginning of this year, some commercial property REITs have declined in the US. However, the REITs of listed IDCs in the US have been recognized by investors and achieved growth against the market (during the first half of 2020, all five US-listed IDC REITs increased from 9 percent to 22 percent).

ViDoRe Benchmark

Qwen2 VL 7B => ColQwen2

PaliGemma => ColPali



Rank	Model	Average	TAT-DQA	Shift Project	Artificial Intelligence	Government Reports	ArxivQA	DocVQA	Healthcare Industry
1	Metric-AI_ColQwen2.5-7b-multi	90.4	80.9	87.7	99.3	95.2	91.7	65.1	97.8
2	Metric-AI_ColQwen2.5-3b-multi	90.3	80.7	88.4	98.9	96.4	92.2	64.4	98.4
3	yydxlv_colqwen2.5-7b-v0.1	90.2	81.9	89.1	98.9	96.2	91.1	63.1	98.5
4	tsystems_colqwen2-7b-v1.0	90.1	78.6	89.3	99.3	96.3	90.7	64.5	99.3
5	Metric-AI_colqwen2.5-3b-multi	90	80.7	86.5	99.4	97.7	93	65.1	98.8
6	yydxlv_colqwen2-7b-v1.0	90	81.7	87.8	100	96.7	89.3	63.1	98.7
7	Alibaba-NLP_gme-Qwen2-VL-7B-I	89.9	76.1	97.3	99.6	99.5	87.6	56.6	99.6
8	Metric-AI_ColQwenStella-2b-mu	89.8	78.7	88.1	99.6	97.1	89.5	62.6	99.6
9	tsystems_colqwen2-2b-v1.0	89.6	79.5	89.9	99.6	95.2	90	60.9	98.9
10	vidore_colqwen2.5-v0.2	89.4	81.1	87.3	99.6	96.4	89.2	63.2	97.9
11	vidore_colqwen2-v1.0	89.3	81.4	90.7	99.4	96.3	88.1	60.6	98.1
12	vidore_colqwen2.5-v0.1	88.8	80.8	85.5	99.3	95.3	88.5	61.9	98.8
13	Alibaba-NLP_gme-Qwen2-VL-2B-I	87.8	71.1	94.3	99	97.9	83.9	54.6	98.9
14	vidore_colqwen2-v0.1	87.3	75.9	86	98.7	92.8	86.1	61.4	98
15	vidore_colsmolvlm-v0.1	86.1	79.5	79.5	98.1	96.9	79.4	60	99.6
16	MrLight_dse-qwen2-2b-mrl-v1	85.8	69.4	82	97.5	96	85.6	57.1	96.4
17	vidore_colpali-v1.3	84.8	70.4	77.4	97.4	96.2	83	58.5	96.9
18	vidore_colpali2-3b-pt-448	84.5	68.6	75.9	98	94.1	82.5	59	97.2
19	vidore_colpali-v1.2	83.9	68	79.1	98.1	94.8	78	57.2	96.7
20	yydxlv_colphi3.5	83.7	73.1	68.2	98.5	95.2	87.2	56.9	95.9

Col-xxx based Multi-Modal RAG

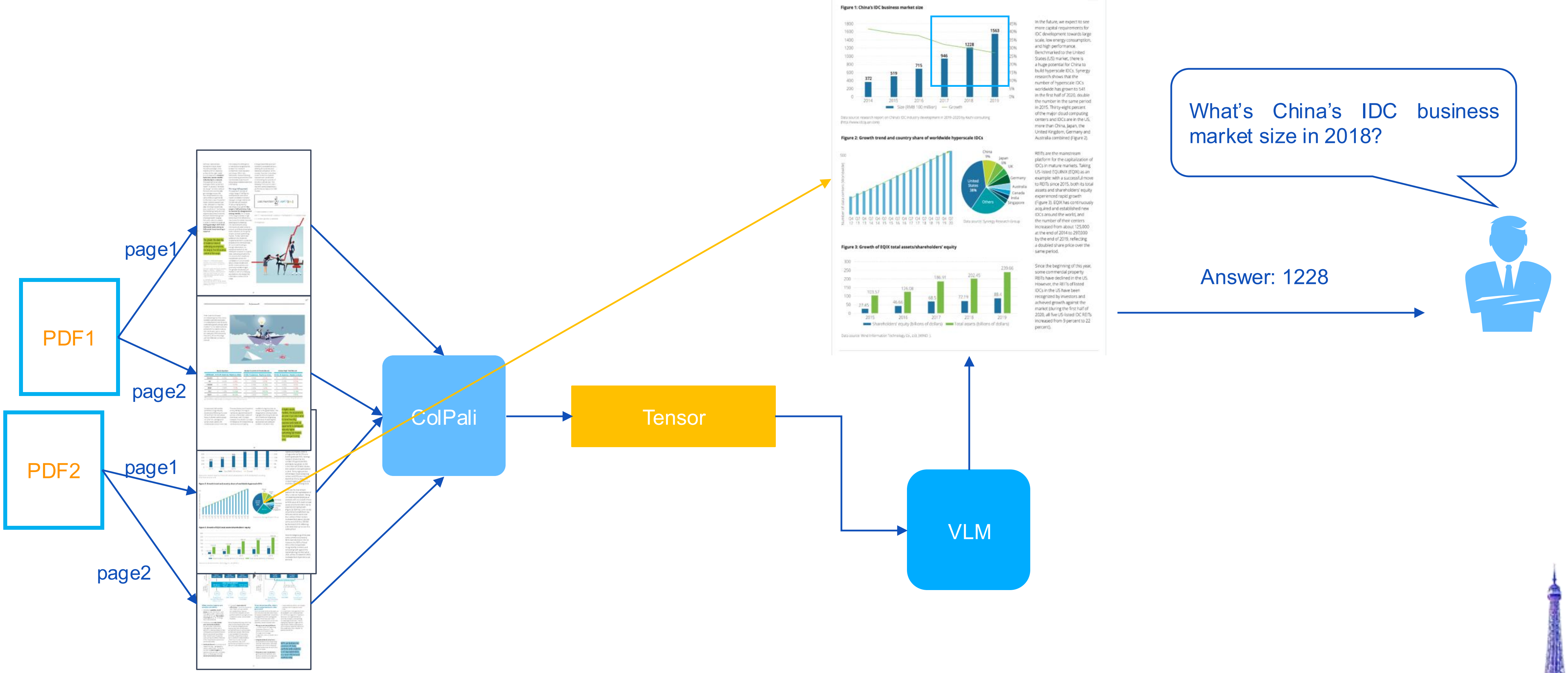


Figure 1: China's IDC business market size

Data source: research report on China's IDC industry development in 2019-2020 by Kezhi consulting (http://www.idcquan.com)

Figure 2: Growth trend and country share of worldwide hyperscale IDCs

Data source: Synergy Research Group

Figure 3: Growth of EQIX total assets/shareholders' equity

Data source: Wind Information Technology Co., Ltd. (WIND)

In the future, we expect to see more capital requirements for IDC development towards large scale, low energy consumption, and high performance. Benchmarked to the United States (US) market, there is a huge potential for China to build hyperscale IDCs. Synergy research shows that the number of hyperscale IDCs worldwide has grown to 541 in the first half of 2020, double the number in the same period in 2015. Thirty-eight percent of the major cloud computing centers and IDCs are in the US, more than China, Japan, the United Kingdom, Germany and Australia combined (Figure 2).

REITs are the mainstream platform for the capitalization of IDCs in mature markets. Taking US-listed EQUINIX (EQIX) as an example: with a successful move to REITs since 2015, both its total assets and shareholders' equity experienced rapid growth (Figure 3). EQIX has continuously acquired and established new IDCs around the world, and the number of their centers increased from about 125,000 at the end of 2014 to 297,000 by the end of 2019, reflecting a doubled share price over the same period.

Since the beginning of this year, some commercial property REITs have declined in the US. However, the REITs of listed IDCs in the US have been recognized by investors and achieved growth against the market (during the first half of 2020, all five US-listed IDC REITs increased from 9 percent to 22 percent).

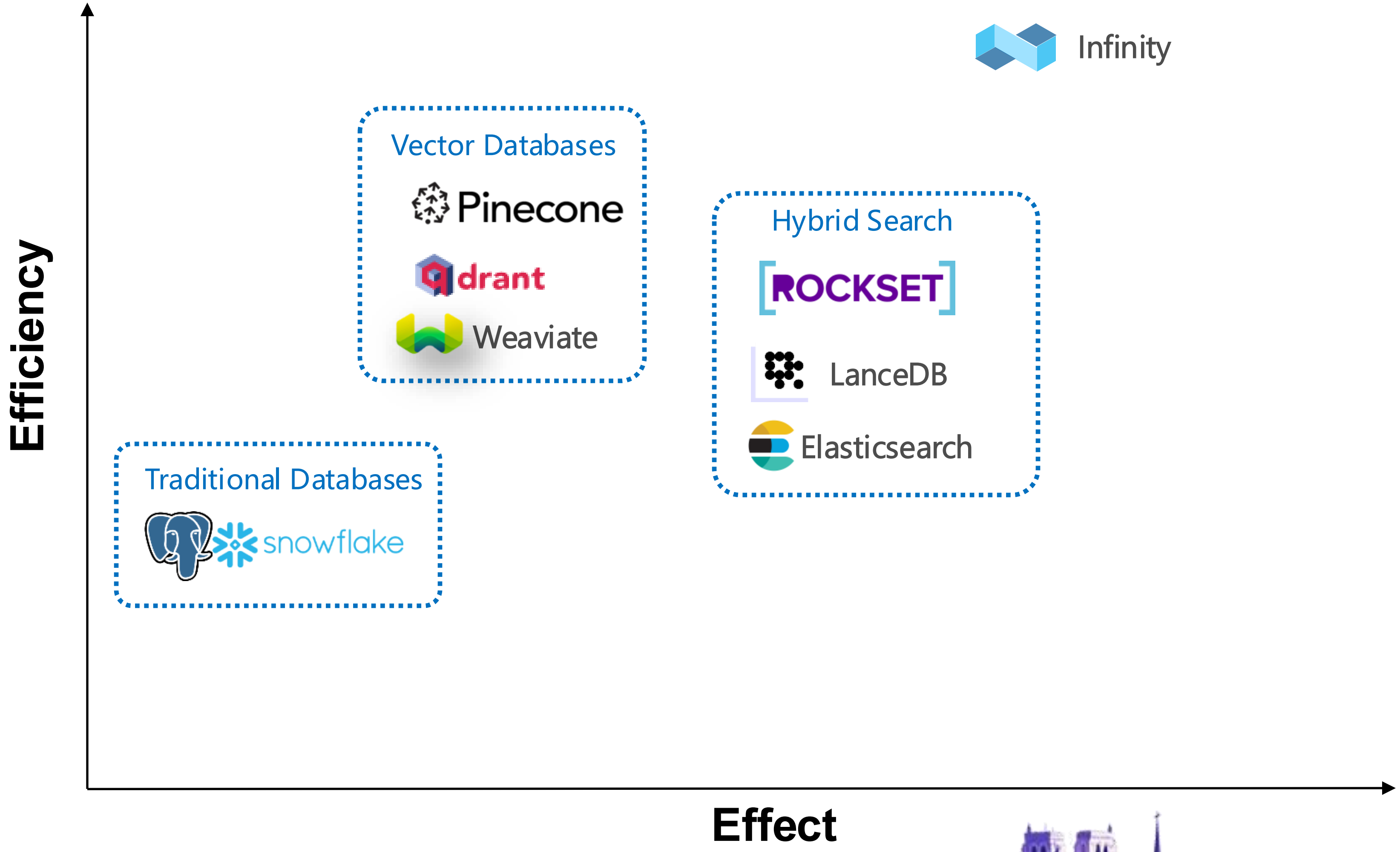
What's China's IDC business market size in 2018?

Answer: 1228

Scaling of Col-xxx Support

- Data base Side
  - Tensor Reranker
  - Binary Quantization
  - Multi-Vector Index
  - Full text index
- Model Side
  - Dimensional Reduction（MRL）
  - Tokens Reduction（ConstBERT/SVD）

Comparison



Efficiency

Effect

Infinity

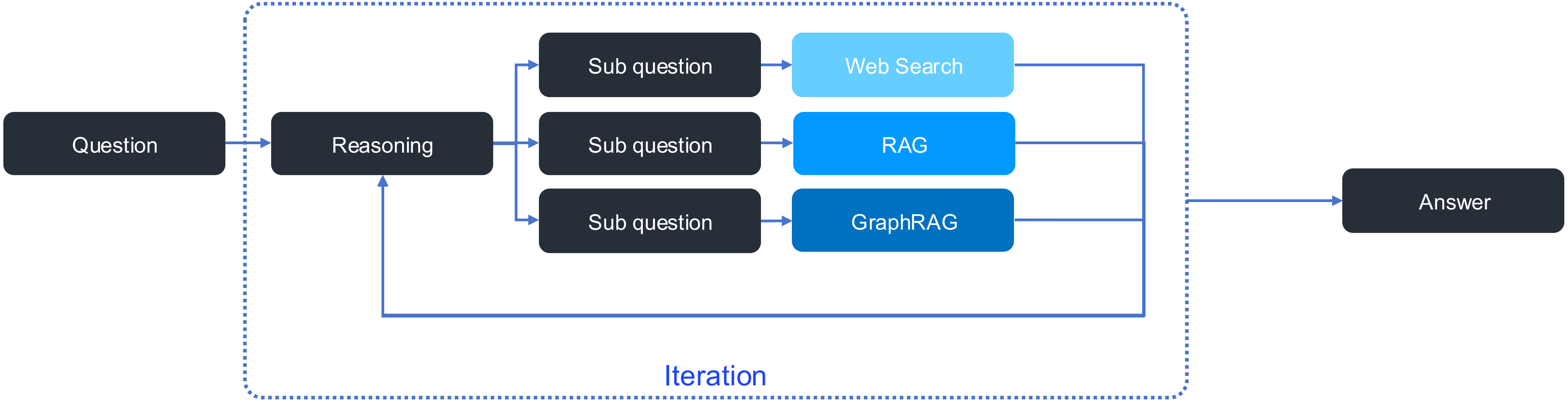
Vector Databases
- Pinecone
- Qdrant
- Weaviate

Hybrid Search
- ROCKSET
- LanceDB
- Elasticsearch

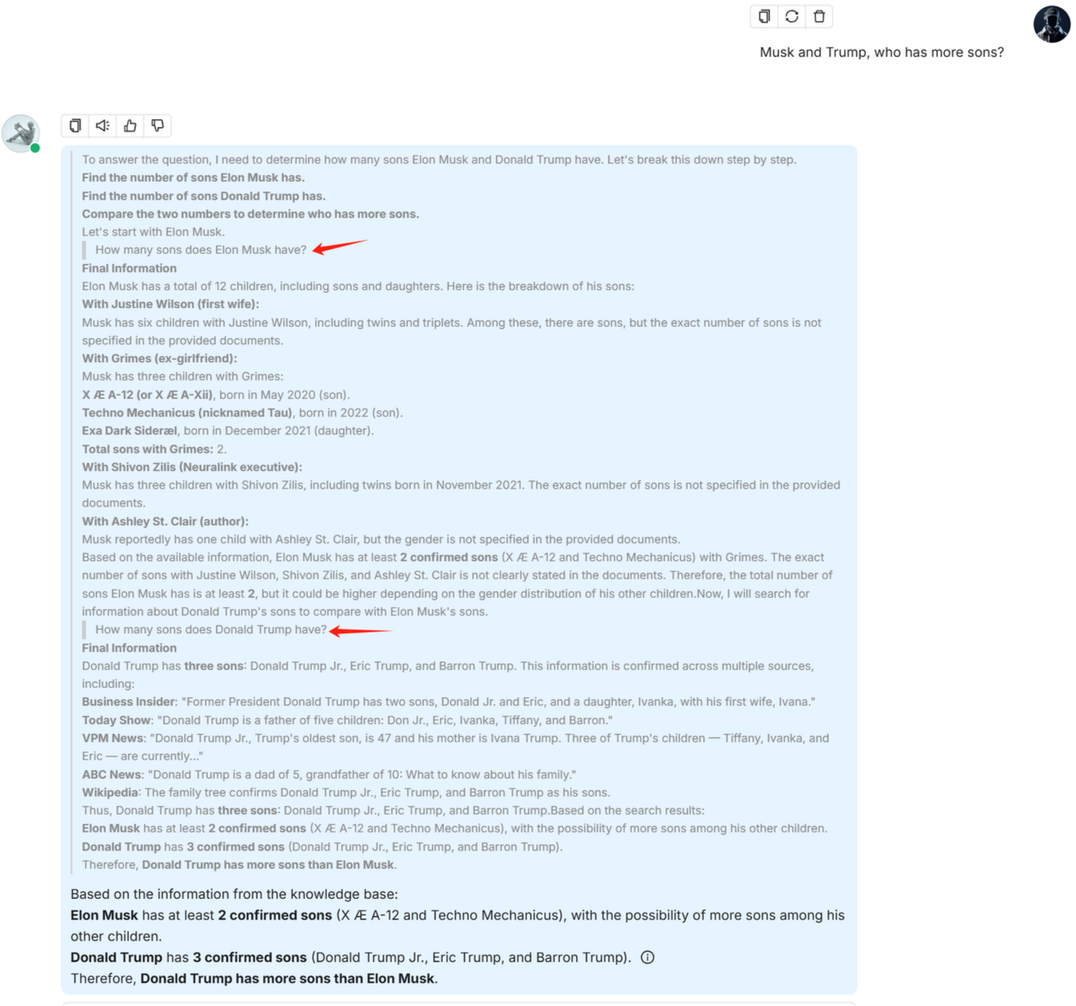
Traditional Databases
- snowflake

Agentic and Future

Agentic Reasoning



Agentic Reasoning



Musk and Trump, who has more sons?

To answer the question, I need to determine how many sons Elon Musk and Donald Trump have. Let's break this down step by step.
Find the number of sons Elon Musk has.
Find the number of sons Donald Trump has.
Compare the two numbers to determine who has more sons.
Let's start with Elon Musk.

> How many sons does Elon Musk have?

Final Information
Elon Musk has a total of 12 children, including sons and daughters. Here is the breakdown of his sons:
With Justine Wilson (first wife):
Musk has six children with Justine Wilson, including twins and triplets. Among these, there are sons, but the exact number of sons is not specified in the provided documents.
With Grimes (ex-girlfriend):
Musk has three children with Grimes:
X Æ A-12 (or X Æ A-Xii), born in May 2020 (son).
Techno Mechanicus (nicknamed Tau), born in 2022 (son).
Exa Dark Sideræl, born in December 2021 (daughter).
Total sons with Grimes: 2.
With Shivon Zilis (Neuralink executive):
Musk has three children with Shivon Zilis, including twins born in November 2021. The exact number of sons is not specified in the provided documents.
With Ashley St. Clair (author):
Musk reportedly has one child with Ashley St. Clair, but the gender is not specified in the provided documents.
Based on the available information, Elon Musk has at least **2 confirmed sons** (X Æ A-12 and Techno Mechanicus) with Grimes. The exact number of sons with Justine Wilson, Shivon Zilis, and Ashley St. Clair is not clearly stated in the documents. Therefore, the total number of sons Elon Musk has is at least **2**, but it could be higher depending on the gender distribution of his other children.Now, I will search for information about Donald Trump's sons to compare with Elon Musk's sons.

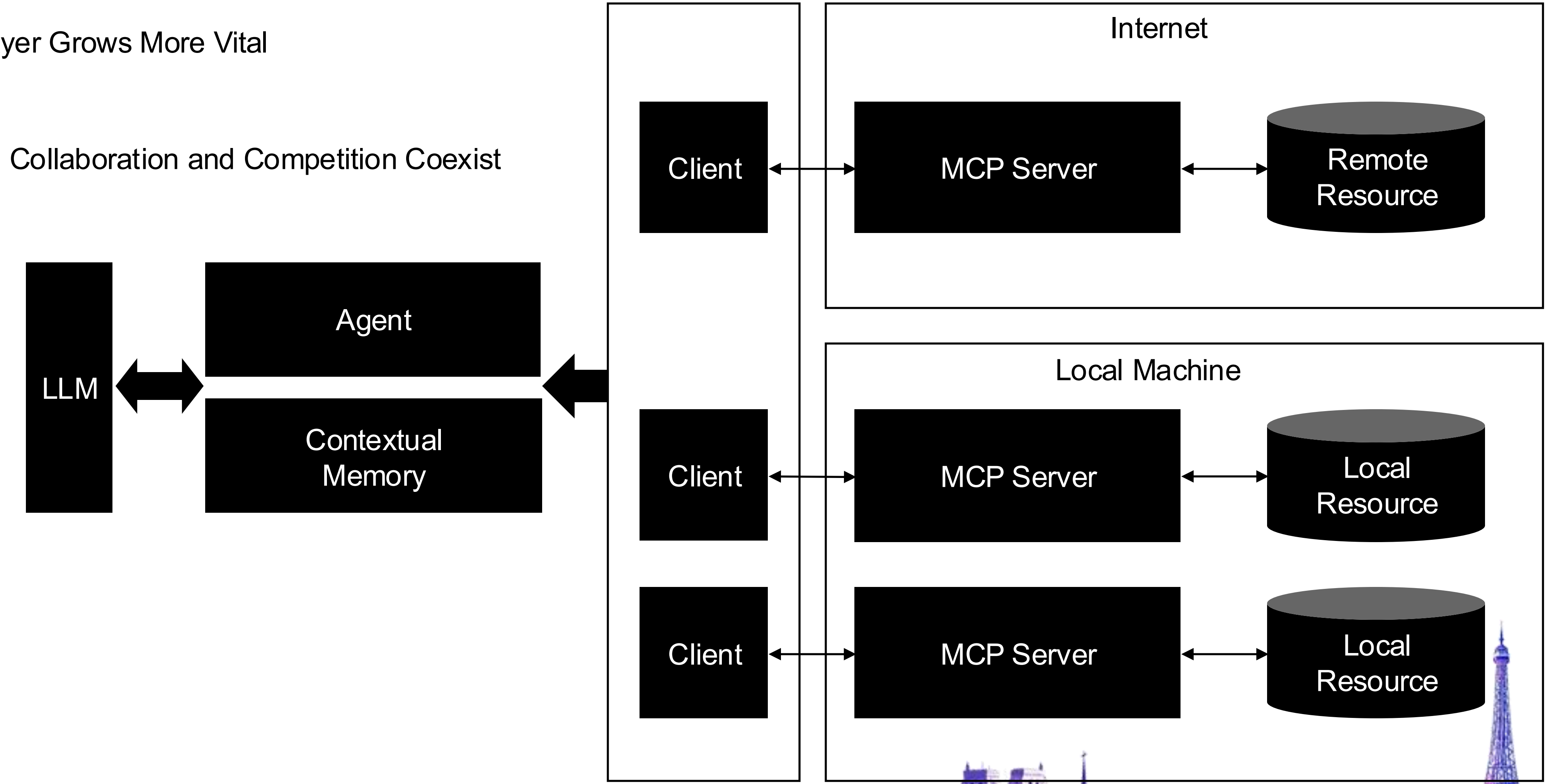
> How many sons does Donald Trump have?

Final Information
Donald Trump has **three sons**: Donald Trump Jr., Eric Trump, and Barron Trump. This information is confirmed across multiple sources, including:
Business Insider: "Former President Donald Trump has two sons, Donald Jr. and Eric, and a daughter, Ivanka, with his first wife, Ivana."
Today Show: "Donald Trump is a father of five children: Don Jr., Eric, Ivanka, Tiffany, and Barron."
VPM News: "Donald Trump Jr., Trump's oldest son, is 47 and his mother is Ivana Trump. Three of Trump's children — Tiffany, Ivanka, and Eric — are currently..."
ABC News: "Donald Trump is a dad of 5, grandfather of 10: What to know about his family."
Wikipedia: The family tree confirms Donald Trump Jr., Eric Trump, and Barron Trump as his sons.
Thus, Donald Trump has **three sons**: Donald Trump Jr., Eric Trump, and Barron Trump.Based on the search results:
Elon Musk has at least **2 confirmed sons** (X Æ A-12 and Techno Mechanicus), with the possibility of more sons among his other children.
Donald Trump has **3 confirmed sons** (Donald Trump Jr., Eric Trump, and Barron Trump).
Therefore, **Donald Trump has more sons than Elon Musk**.

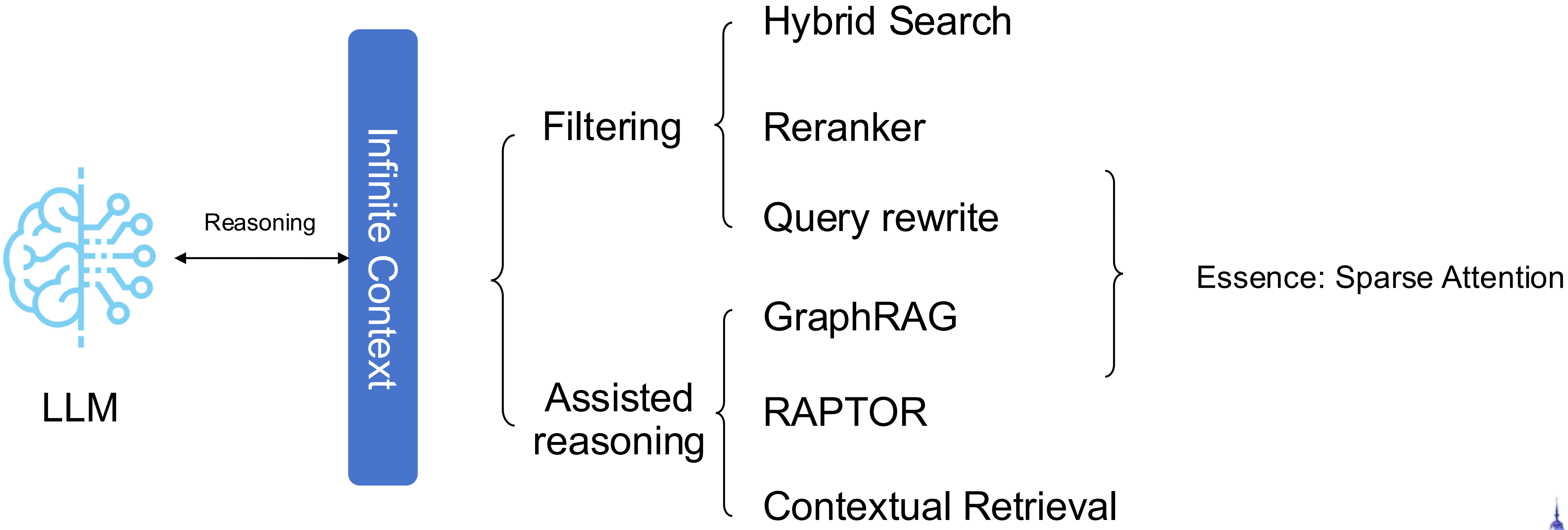
Based on the information from the knowledge base:
Elon Musk has at least **2 confirmed sons** (X Æ A-12 and Techno Mechanicus), with the possibility of more sons among his other children.
Donald Trump has **3 confirmed sons** (Donald Trump Jr., Eric Trump, and Barron Trump).
Therefore, **Donald Trump has more sons than Elon Musk**.

RAG ⇔ Agent

- RAG As a Distinct Layer Grows More Vital
- RAG & Agent:
  Frequent Interaction, Collaboration and Competition Coexist



The Essence of Memory



LLM

Reasoning

Infinite Context

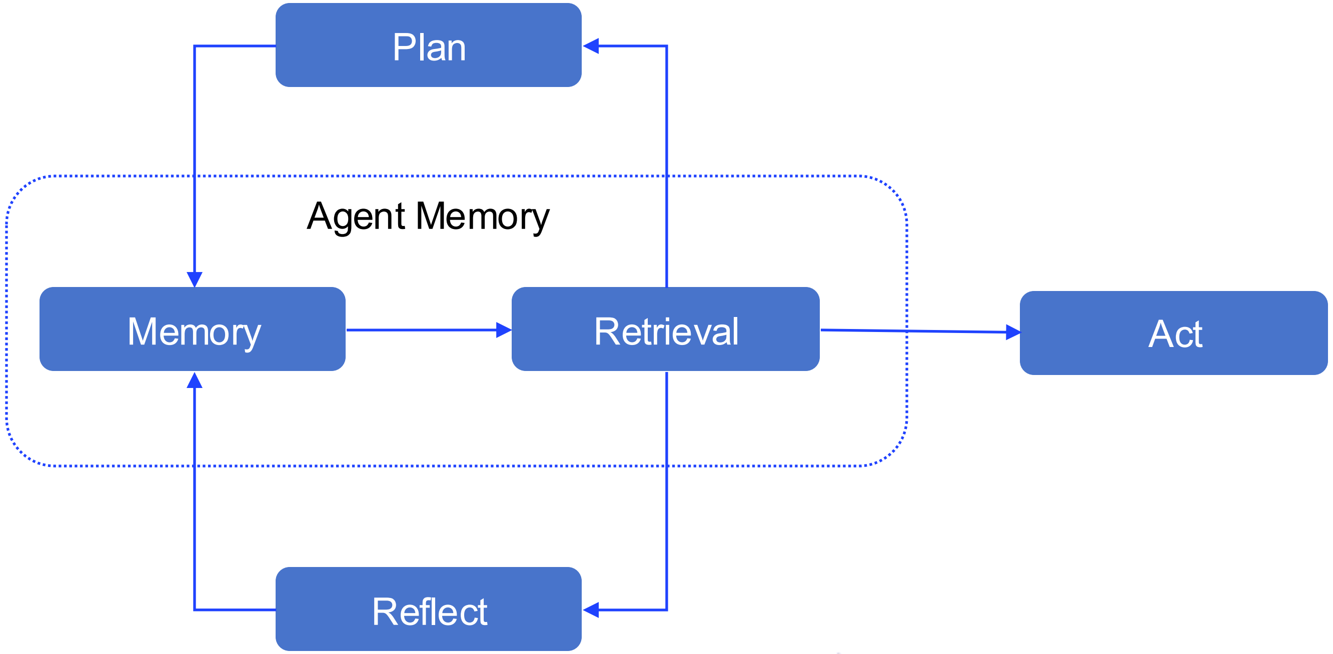
Filtering
- Hybrid Search
- Reranker
- Query rewrite

Assisted reasoning
- GraphRAG
- RAPTOR
- Contextual Retrieval

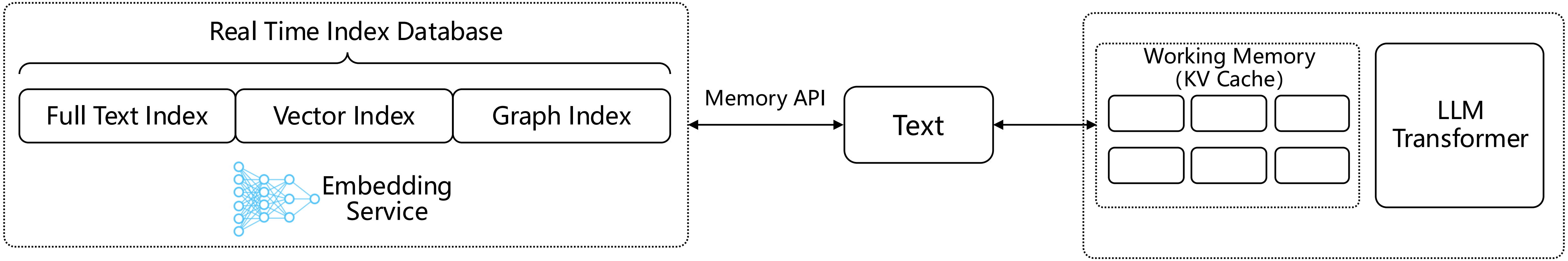
Essence: Sparse Attention

Agent Memory = RAG ?

- Retrieval Differences—Temporal/Weights
- Memory Decay
- Agent Facilities



RAG/Memory Evolution



Real Time Index Database

Full Text Index

Vector Index

Graph Index

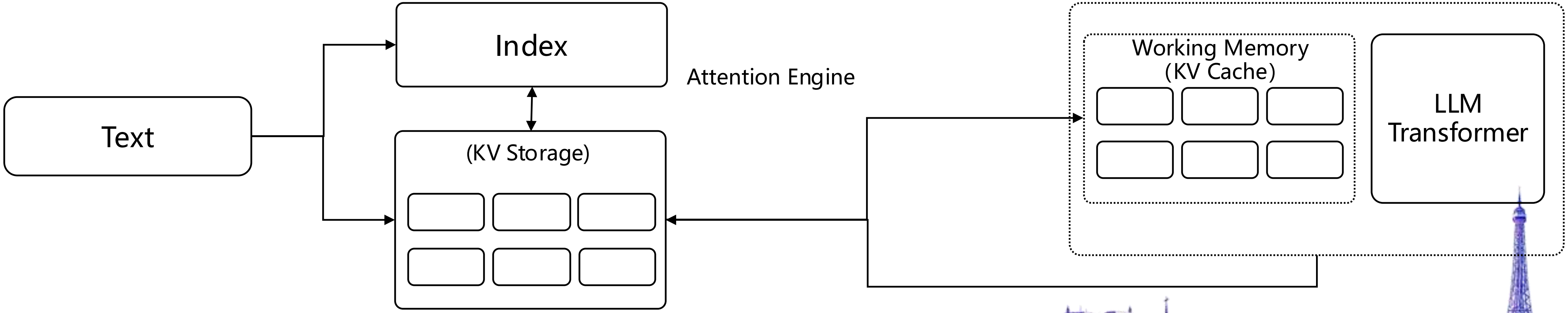
Embedding Service

Memory API

Text

Working Memory (KV Cache)

LLM Transformer



Text

Index

(KV Storage)

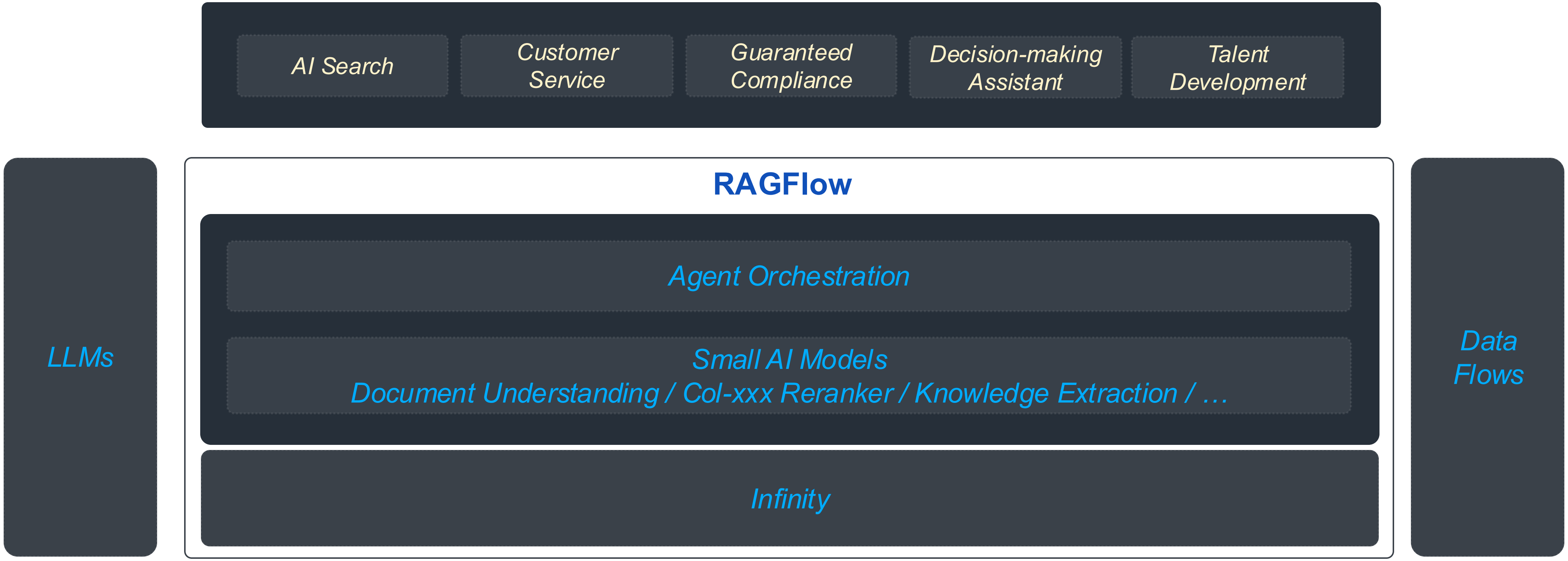
Attention Engine

Working Memory (KV Cache)

LLM Transformer

Summary

RAGFlow Ecosystem



AI Search

Customer Service

Guaranteed Compliance

Decision-making Assistant

Talent Development

LLMs

RAGFlow

Agent Orchestration

Small AI Models
Document Understanding / Col-xxx Reranker / Knowledge Extraction / …

Infinity

Data Flows

RAG is not LLMOps, RAG is the database of LLM era

THANK YOU

https://github.com/infiniflow/ragflow
https://github.com/infiniflow/infinity