BERLIN BUZZWORDS

15-17 June 2025

Contexts & Machines: How Document Parsing Shapes RAG results

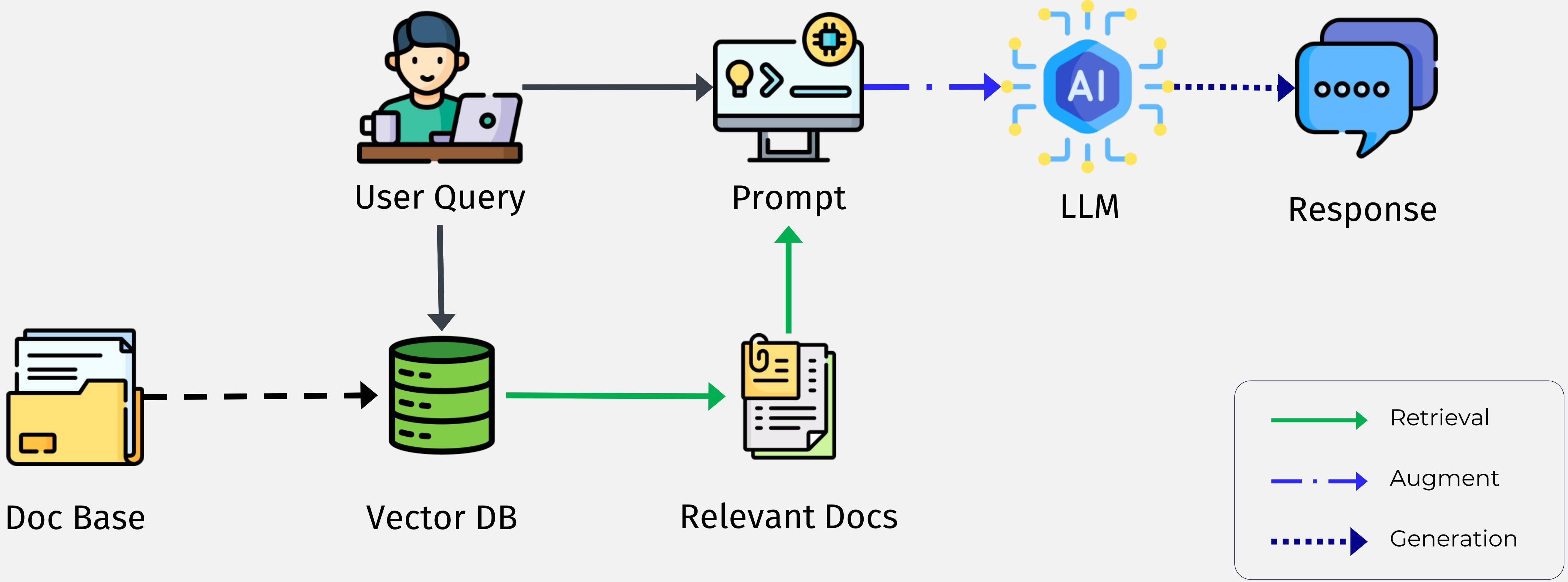
Alessio Vertemati

Andrea Ponti

oneofftech.xyz

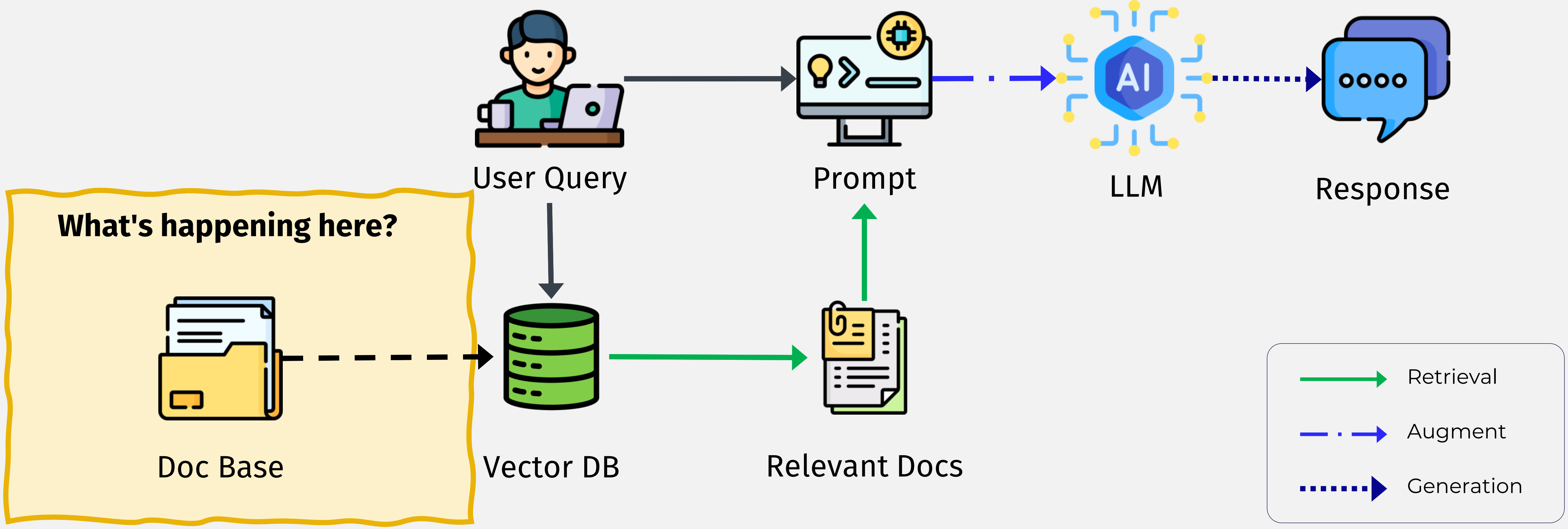
Retrieval Augmented Generation (RAG)

RAG frameworks enhance large language models by providing external knowledge to ground their responses.



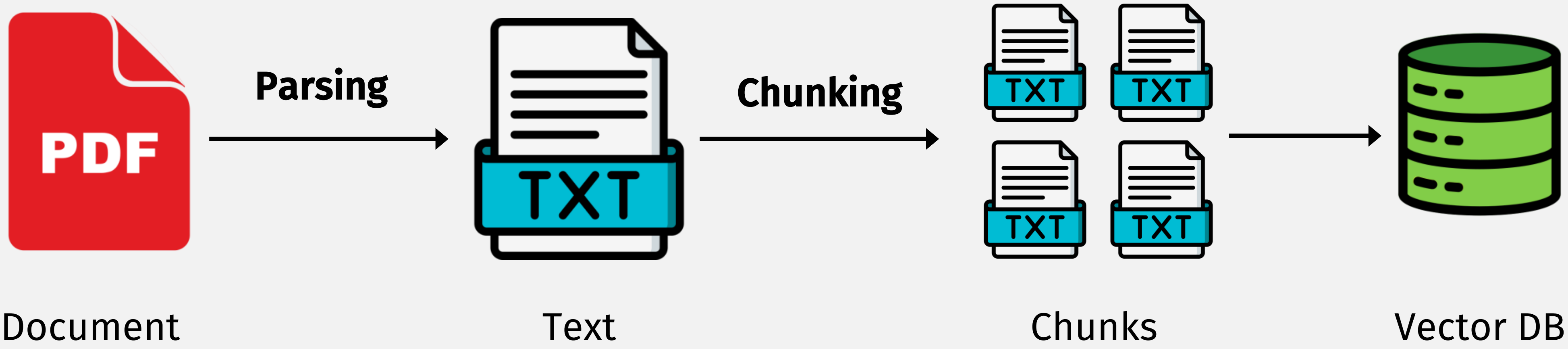
Retrieval Augmented Generation (RAG)

RAG frameworks enhance large language models by providing external knowledge to ground their responses.



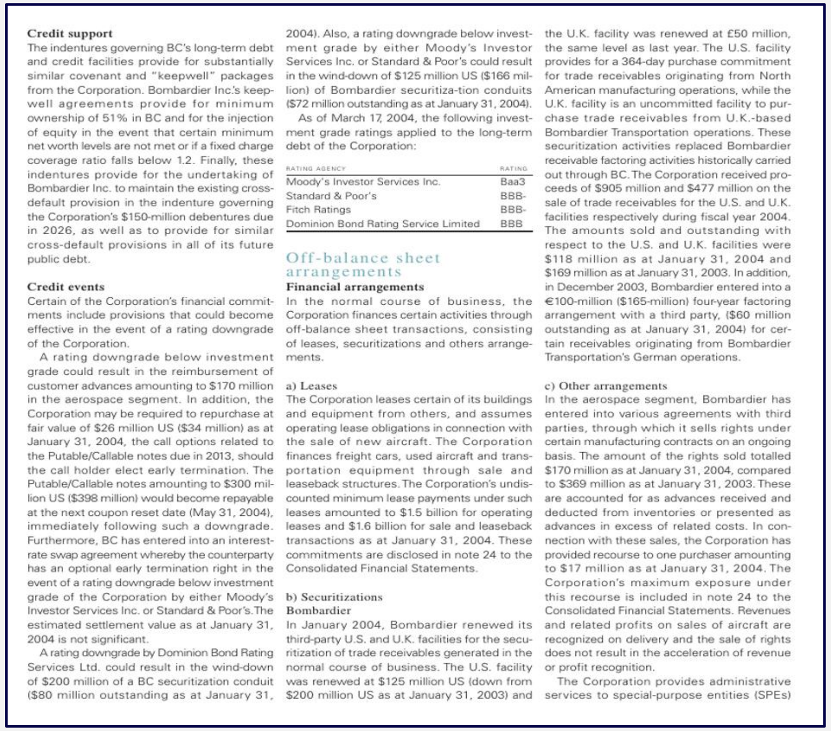
Document Processing for RAG

Parsing and chunking define what the retriever sees — and if it sees the wrong thing, the generator fails.



Why is it so hard?

Common Challenges: Reading Order, Table Parsing, Headers, Footers, …



Credit support

The indentures governing BC's long-term debt and credit facilities provide for substantially similar covenant and "keepwell" packages from the Corporation. Bombardier Inc.'s keepwell agreements provide for minimum ownership of 51% in BC and for the injection of equity in the event that certain minimum net worth levels are not met or if a fixed charge coverage ratio falls below 1.2. Finally, these indentures provide for the undertaking of Bombardier Inc. to maintain the existing cross-default provision in the indenture governing the Corporation's $150-million debentures due in 2026, as well as to provide for similar cross-default provisions in all of its future public debt.

Credit events

Certain of the Corporation's financial commitments include provisions that could become effective in the event of a rating downgrade of the Corporation.

A rating downgrade below investment grade could result in the reimbursement of customer advances amounting to $170 million in the aerospace segment. In addition, the Corporation may be required to repurchase at fair value of $26 million US ($34 million) as at January 31, 2004, the call options related to the Putable/Callable notes due in 2013, should the call holder elect early termination. The Putable/Callable notes amounting to $300 million US ($398 million) would become repayable at the next coupon reset date (May 31, 2004), immediately following such a downgrade. Furthermore, BC has entered into an interest-rate swap agreement whereby the counterparty has an optional early termination right in the event of a rating downgrade below investment grade of the Corporation by either Moody's Investor Services Inc. or Standard & Poor's.The estimated settlement value as at January 31, 2004 is not significant.

A rating downgrade by Dominion Bond Rating Services Ltd. could result in the wind-down of $200 million of a BC securitization conduit ($80 million outstanding as at January 31, 2004). Also, a rating downgrade below investment grade by either Moody's Investor Services Inc. or Standard & Poor's could result in the wind-down of $125 million US ($166 million) of Bombardier securitiza-tion conduits ($72 million outstanding as at January 31, 2004).

As of March 17, 2004, the following investment grade ratings applied to the long-term debt of the Corporation:

RATING AGENCY	RATING
Moody's Investor Services Inc.	Baa3
Standard & Poor's	BBB-
Fitch Ratings	BBB-
Dominion Bond Rating Service Limited	BBB

Off-balance sheet arrangements

Financial arrangements

In the normal course of business, the Corporation finances certain activities through off-balance sheet transactions, consisting of leases, securitizations and others arrangements.

a) Leases

The Corporation leases certain of its buildings and equipment from others, and assumes operating lease obligations in connection with the sale of new aircraft. The Corporation finances freight cars, used aircraft and transportation equipment through sale and leaseback structures. The Corporation's undiscounted minimum lease payments under such leases amounted to $1.5 billion for operating leases and $1.6 billion for sale and leaseback transactions as at January 31, 2004. These commitments are disclosed in note 24 to the Consolidated Financial Statements.

b) Securitizations

Bombardier

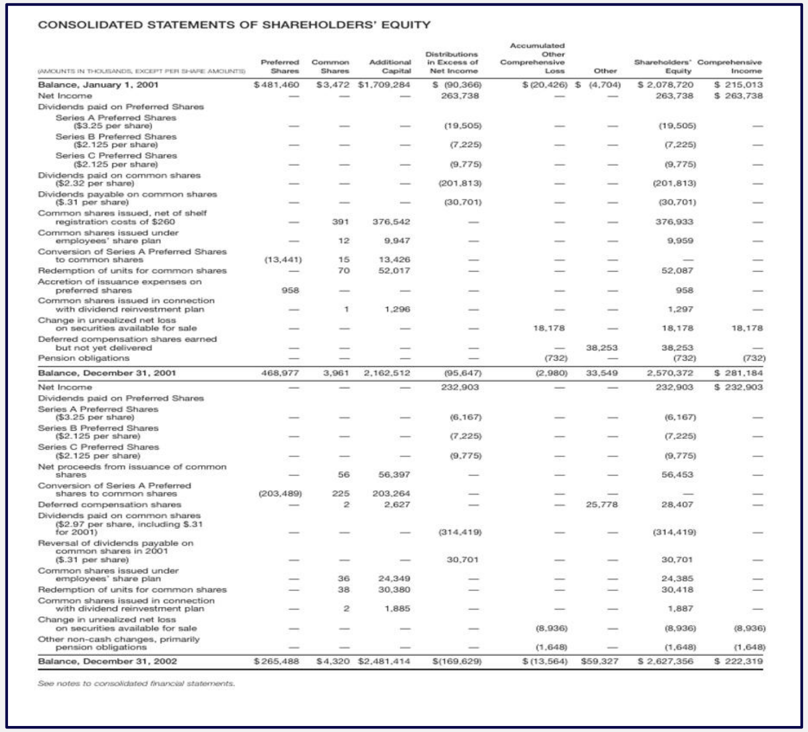
In January 2004, Bombardier renewed its third-party U.S. and U.K. facilities for the securitization of trade receivables generated in the normal course of business. The U.S. facility was renewed at $125 million US (down from $200 million US as at January 31, 2003) and the U.K. facility was renewed at £50 million, the same level as last year. The U.S. facility provides for a 364-day purchase commitment for trade receivables originating from North American manufacturing operations, while the U.K. facility is an uncommitted facility to purchase trade receivables from U.K.-based Bombardier Transportation operations. These securitization activities replaced Bombardier receivable factoring activities historically carried out through BC. The Corporation received proceeds of $905 million and $477 million on the sale of trade receivables for the U.S. and U.K. facilities respectively during fiscal year 2004. The amounts sold and outstanding with respect to the U.S. and U.K. facilities were $118 million as at January 31, 2004 and $169 million as at January 31, 2003. In addition, in December 2003, Bombardier entered into a €100-million ($165-million) four-year factoring arrangement with a third party, ($60 million outstanding as at January 31, 2004) for certain receivables originating from Bombardier Transportation's German operations.

c) Other arrangements

In the aerospace segment, Bombardier has entered into various agreements with third parties, through which it sells rights under certain manufacturing contracts on an ongoing basis. The amount of the rights sold totalled $170 million as at January 31, 2004, compared to $369 million as at January 31, 2003. These are accounted for as advances received and deducted from inventories or presented as advances in excess of related costs. In connection with these sales, the Corporation has provided recourse to one purchaser amounting to $17 million as at January 31, 2004. The Corporation's maximum exposure under this recourse is included in note 24 to the Consolidated Financial Statements. Revenues and related profits on sales of aircraft are recognized on delivery and the sale of rights does not result in the acceleration of revenue or profit recognition.

The Corporation provides administrative services to special-purpose entities (SPEs)

Multi-column documents

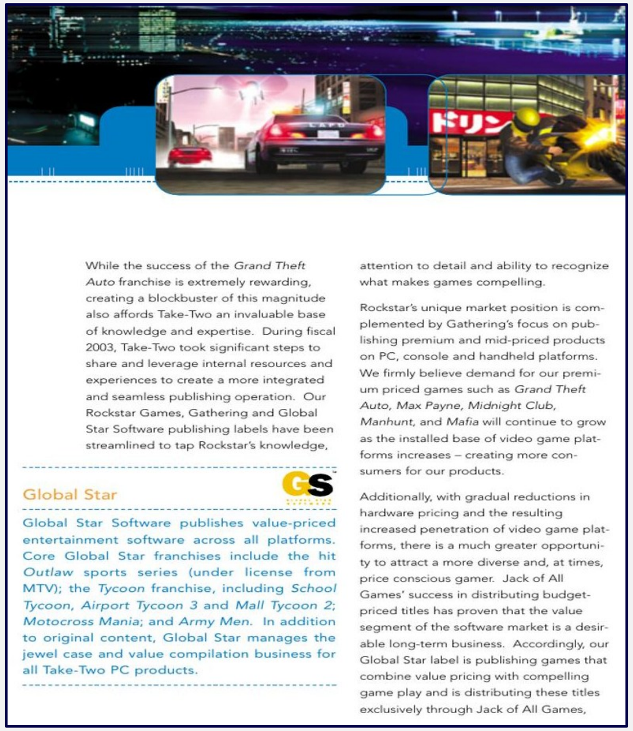


CONSOLIDATED STATEMENTS OF SHAREHOLDERS' EQUITY

(AMOUNTS IN THOUSANDS, EXCEPT PER SHARE AMOUNTS)	Preferred Shares	Common Shares	Additional Capital	Distributions in Excess of Net Income	Accumulated Other Comprehensive Loss	Other	Shareholders' Equity	Comprehensive Income
Balance, January 1, 2001	$481,460	$3,472	$1,709,284	$ (90,366)	$ (20,426)	$ (4,704)	$ 2,078,720	$ 215,013
Net Income	—	—	—	263,738	—	—	263,738	$ 263,738
Dividends paid on Preferred Shares								
Series A Preferred Shares ($3.25 per share)	—	—	—	(19,505)	—	—	(19,505)	—
Series B Preferred Shares ($2.125 per share)	—	—	—	(7,225)	—	—	(7,225)	—
Series C Preferred Shares ($2.125 per share)	—	—	—	(9,775)	—	—	(9,775)	—
Dividends paid on common shares ($2.32 per share)	—	—	—	(201,813)	—	—	(201,813)	—
Dividends payable on common shares ($.31 per share)	—	—	—	(30,701)	—	—	(30,701)	—
Common shares issued, net of shelf registration costs of $260	—	391	376,542	—	—	—	376,933	—
Common shares issued under employees' share plan	—	12	9,947	—	—	—	9,959	—
Conversion of Series A Preferred Shares to common shares	(13,441)	15	13,426	—	—	—	—	—
Redemption of units for common shares	—	70	52,017	—	—	—	52,087	—
Accretion of issuance expenses on preferred shares	958	—	—	—	—	—	958	—
Common shares issued in connection with dividend reinvestment plan	—	1	1,296	—	—	—	1,297	—
Change in unrealized net loss on securities available for sale	—	—	—	—	18,178	—	18,178	18,178
Deferred compensation shares earned but not yet delivered	—	—	—	—	—	38,253	38,253	—
Pension obligations	—	—	—	—	(732)	—	(732)	(732)
Balance, December 31, 2001	468,977	3,961	2,162,512	(95,647)	(2,980)	33,549	2,570,372	$ 281,184
Net Income	—	—	—	232,903	—	—	232,903	$ 232,903
Dividends paid on Preferred Shares								
Series A Preferred Shares ($3.25 per share)	—	—	—	(6,167)	—	—	(6,167)	—
Series B Preferred Shares ($2.125 per share)	—	—	—	(7,225)	—	—	(7,225)	—
Series C Preferred Shares ($2.125 per share)	—	—	—	(9,775)	—	—	(9,775)	—
Net proceeds from issuance of common shares	—	56	56,397	—	—	—	56,453	—
Conversion of Series A Preferred shares to common shares	(203,489)	225	203,264	—	—	—	—	—
Deferred compensation shares	—	2	2,627	—	—	25,778	28,407	—
Dividends paid on common shares ($2.97 per share, including $.31 for 2001)	—	—	—	(314,419)	—	—	(314,419)	—
Reversal of dividends payable on common shares in 2001 ($.31 per share)	—	—	—	30,701	—	—	30,701	—
Common shares issued under employees' share plan	—	36	24,349	—	—	—	24,385	—
Redemption of units for common shares	—	38	30,380	—	—	—	30,418	—
Common shares issued in connection with dividend reinvestment plan	—	2	1,885	—	—	—	1,887	—
Change in unrealized net loss on securities available for sale	—	—	—	—	(8,936)	—	(8,936)	(8,936)
Other non-cash changes, primarily pension obligations	—	—	—	—	(1,648)	—	(1,648)	(1,648)
Balance, December 31, 2002	$265,488	$4,320	$2,481,414	$(169,629)	$ (13,564)	$59,327	$ 2,627,356	$ 222,319

See notes to consolidated financial statements.

Complex Tables



While the success of the *Grand Theft Auto* franchise is extremely rewarding, creating a blockbuster of this magnitude also affords Take-Two an invaluable base of knowledge and expertise. During fiscal 2003, Take-Two took significant steps to share and leverage internal resources and experiences to create a more integrated and seamless publishing operation. Our Rockstar Games, Gathering and Global Star Software publishing labels have been streamlined to tap Rockstar's knowledge, attention to detail and ability to recognize what makes games compelling.

Rockstar's unique market position is complemented by Gathering's focus on publishing premium and mid-priced products on PC, console and handheld platforms. We firmly believe demand for our premium priced games such as *Grand Theft Auto, Max Payne, Midnight Club, Manhunt*, and *Mafia* will continue to grow as the installed base of video game platforms increases – creating more consumers for our products.

Additionally, with gradual reductions in hardware pricing and the resulting increased penetration of video game platforms, there is a much greater opportunity to attract a more diverse and, at times, price conscious gamer. Jack of All Games' success in distributing budget-priced titles has proven that the value segment of the software market is a desirable long-term business. Accordingly, our Global Star label is publishing games that combine value pricing with compelling game play and is distributing these titles exclusively through Jack of All Games,

Global Star

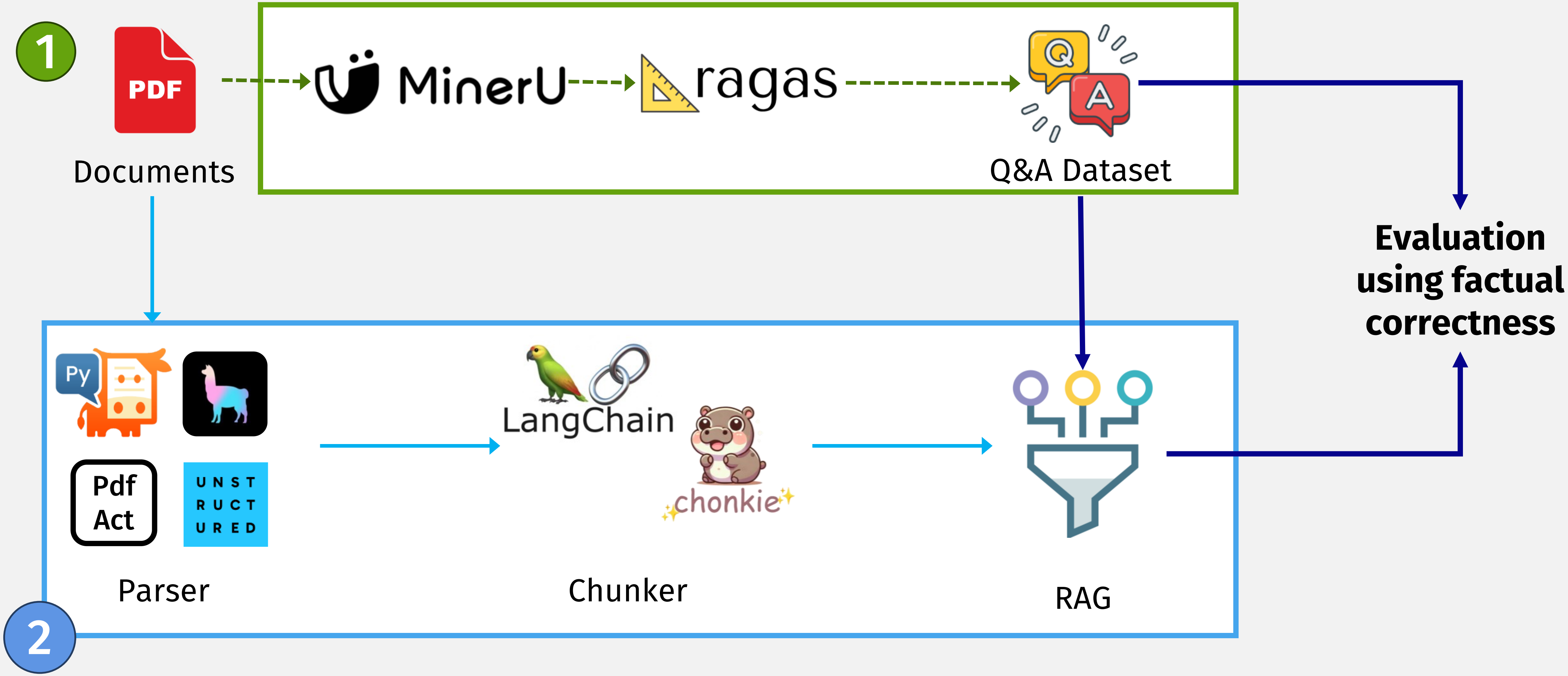
Global Star Software publishes value-priced entertainment software across all platforms. Core Global Star franchises include the hit *Outlaw* sports series (under license from MTV); the *Tycoon* franchise, including *School Tycoon, Airport Tycoon 3* and *Mall Tycoon 2*; *Motocross Mania*; and *Army Men*. In addition to original content, Global Star manages the jewel case and value compilation business for all Take-Two PC products.

Complex Layout

Methodology

1 Dataset Building

2 Evaluation



1

Documents

Q&A Dataset

Evaluation using factual correctness

2

Parser

Chunker

RAG

>BLN
BZZ/
WRDS

Document Processing in RAG

How the document processing strategy impact the performance of RAG

Our setup

4 parsers
- PyMuPDF
- PdfAct
- LlamaParse
- Unstructured

2 chunkers
- LangChain – *Rule-Based*
- Chonkie - *Semantic*

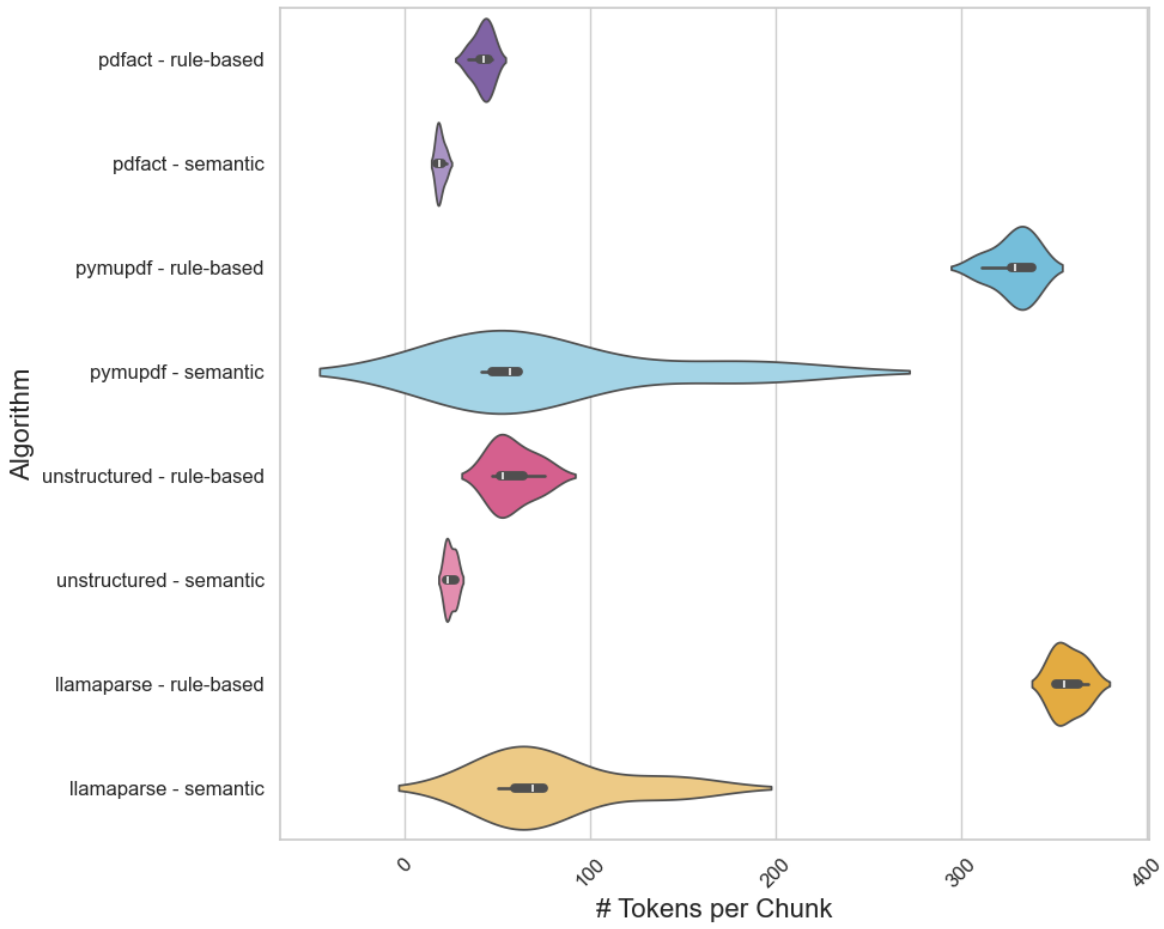
5 documents
~100 pages each
Projects' report

255 questions
single-hop specific
multi-hop specific
multi-hop abstract

Evaluation
RAGAS Factual Correctness

Chunk Size Effect

Chunk Size Distribution by parser



- **Rule-Based chunking** generally creates bigger chunks.
- **Semantic chunking** have higher variance of chunk size.
- ***PdfAct*** (open-source) has a behaviour similar to ***unstructured*** (cloud hosted)
- ***PyMuPDF*** (open-source) has a behaviour similar to ***LlamaParse*** (closed-source)

Results

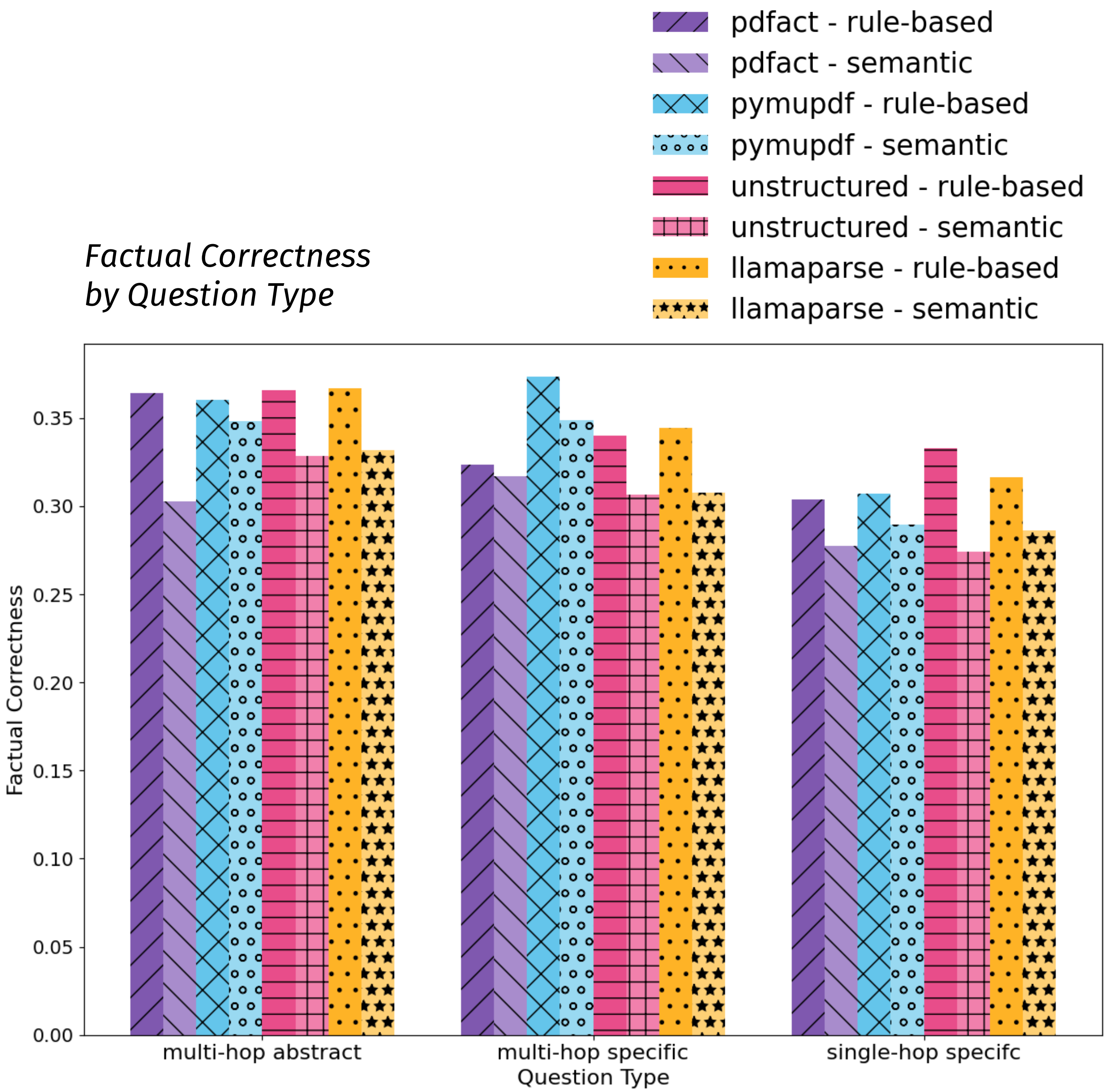
Multi-Hop Abstract:
Chunk size invariance

Multi-Hop Specific:
Larger chunks

Single-Hop Specific:
Shorter chunks

Complexity

Rule-based chunking works great!



Factual Correctness by Question Type

>BLN
BZZ/
WRDS

The Impact of Document Type

How to choose the right document parser?

parxy https://github.com/OneOffTech/parxy

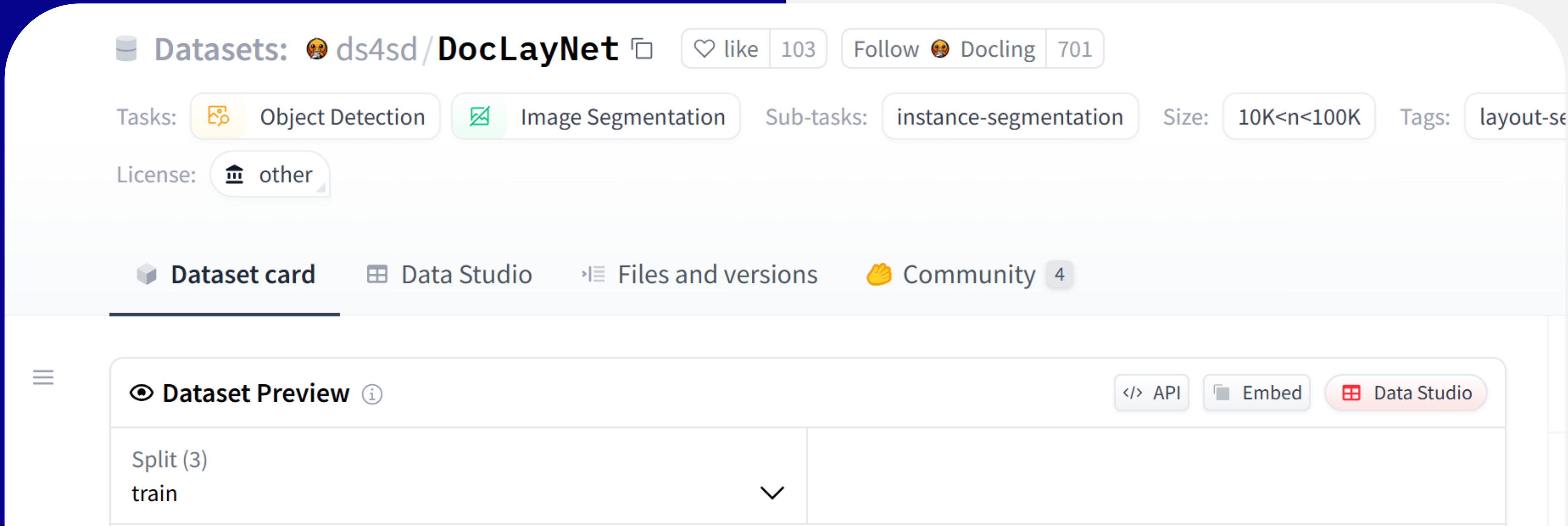
>BLN
BZZ/
WRDS

Our setup

4 parsers

- PyMuPDF
- PdfAct
- LlamaParse
- Unstructured

DocLayNet Dataset
80k pages
6 categories
11 classes
https://huggingface.co/datasets/ds4sd/DocLayNet



Datasets: ds4sd/DocLayNet like 103 Follow Docling 701

Tasks: Object Detection Image Segmentation Sub-tasks: instance-segmentation Size: 10K<n<100K Tags: layout-se

License: other

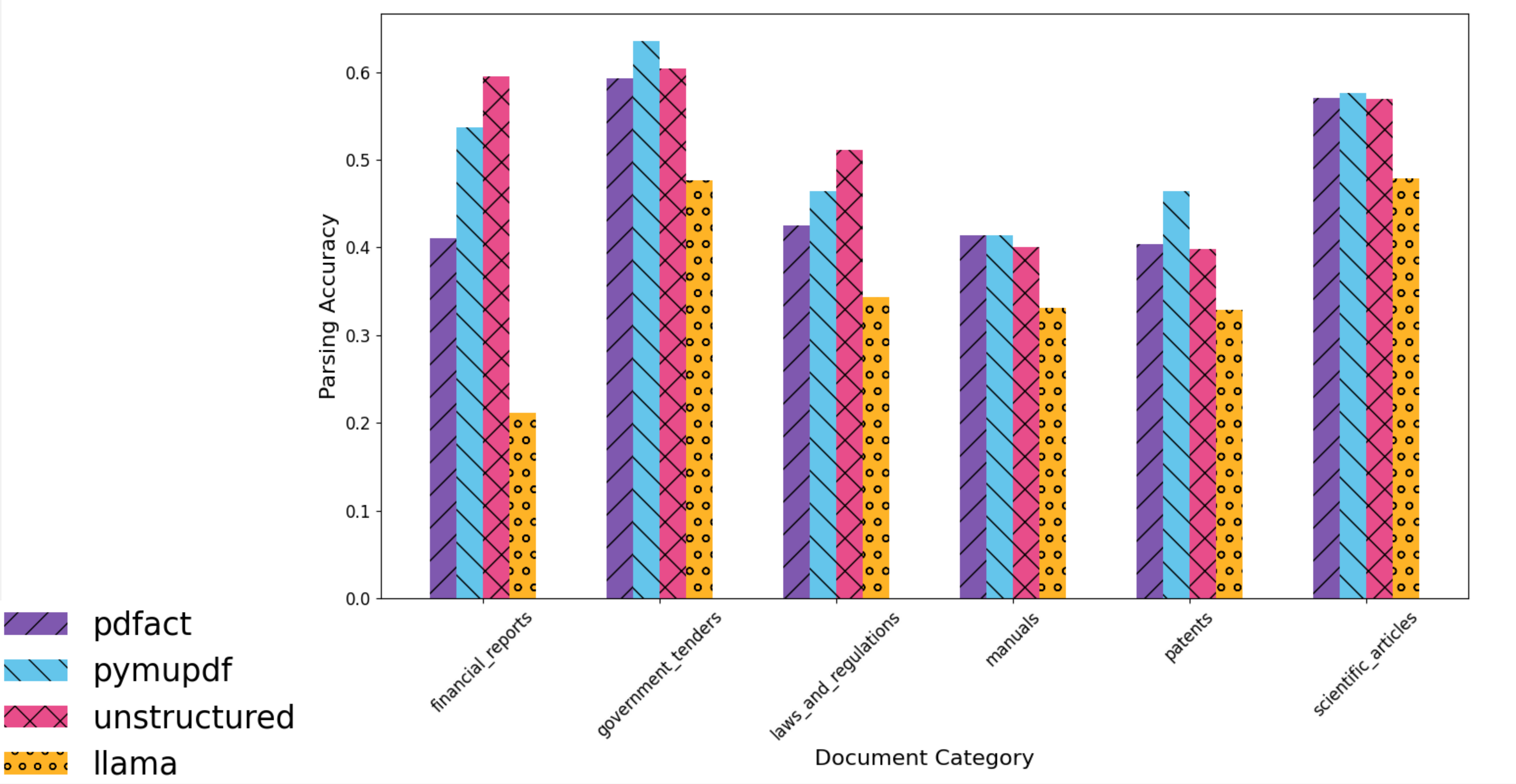
Dataset card | Data Studio | Files and versions | Community 4

Dataset Preview API Embed Data Studio

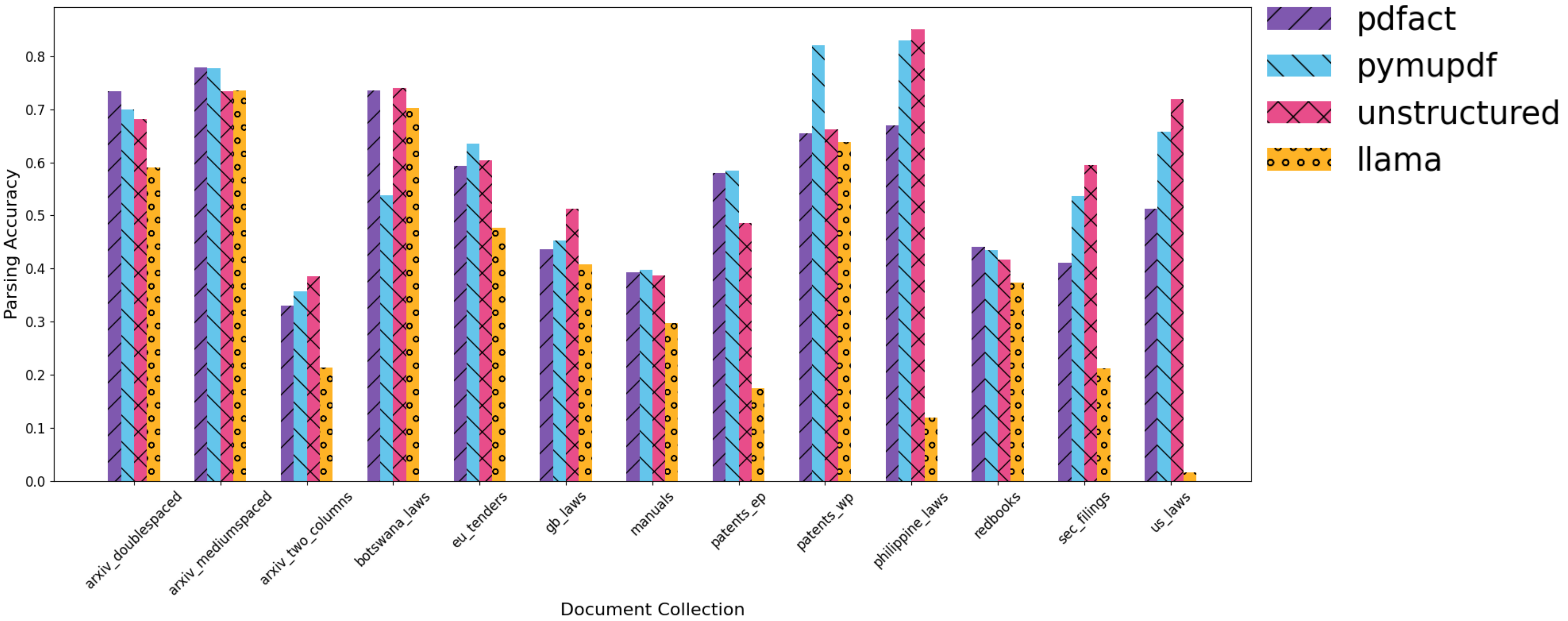
Split (3)
train

Parser vs Document category

Accuracy calculated using Text Similarity Ratio between the extracted text and the ground truth



How To Choose the Right Parser?



Key take-aways

Everything depend on the **document** and **question** **types**

- Document type (and structure) influence the parser choice
- Question type influence chunking strategy

If your RAG system isn't performing well **look first at what you're retrieving**—and how that content is processed!

Some links

- https://github.com/data-house/pdfact
- https://unstructured.io/
- https://github.com/oneofftech/awesome-pdf
- https://github.com/opendatalab/MinerU
- https://parxy.eu
- https://docs.cloud.llamaindex.ai/llamaparse/getting_started

ONEOFF

Check our blog oneofftech.xyz/blog

http://www.oneofftech.xyz/

>BLN
BZZ/
WRDS

Icon Credits

- Scissors icons created by Gulraiz - Flaticon
- Parsing icons created by Good Ware - Flaticon
- Ai technology icons created by FACH - Flaticon
- Embedded icons created by Freepik - Flaticon
- Database icons created by Creatype - Flaticon
- Command icons created by Freepik - Flaticon
- Message icons created by Freepik - Flaticon
- Document icons created by Freepik - Flaticon
- Question icons created by Flat-icons-com - Flaticon
- Screening icons created by Vectors Tank - Flaticon
- Computer icons created by Freepik - Flaticon
- Pdf icons created by egorpolyakov - Flaticon
- Multimedia icons created by surang - Flaticon