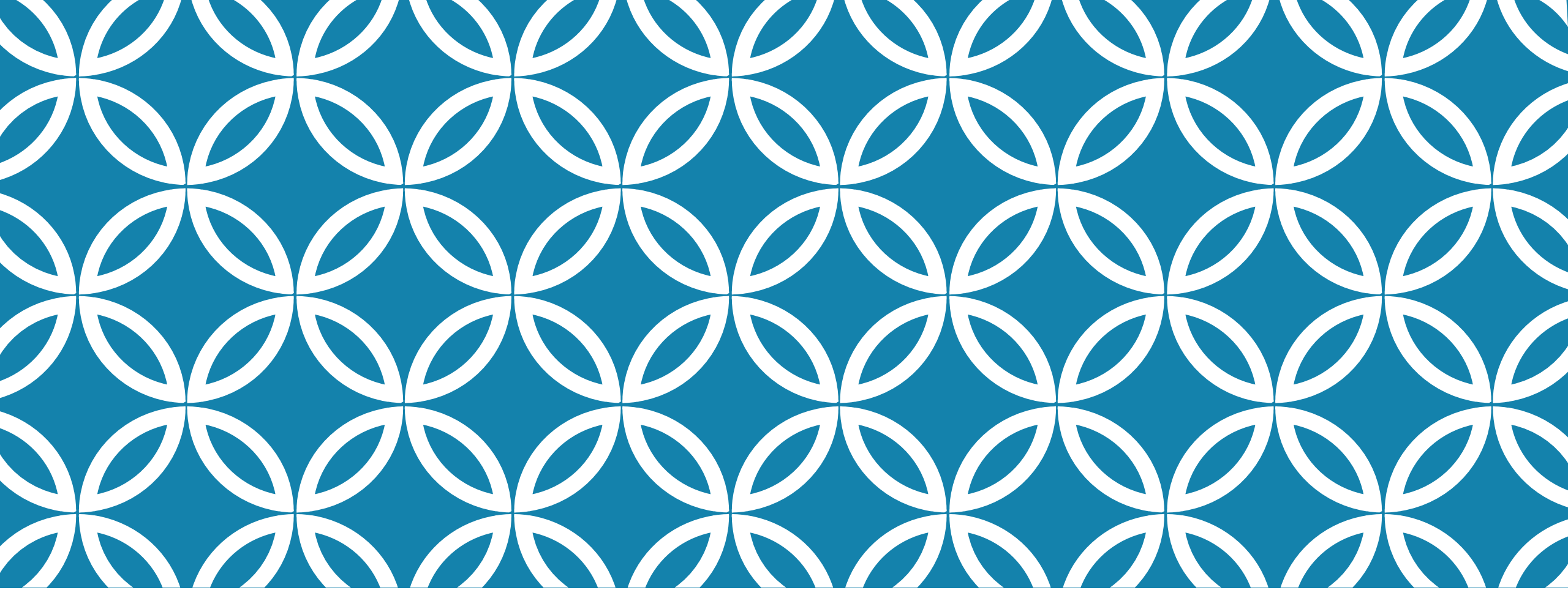


PARSING PDF FILE AND RETRIEVING INFORMATION

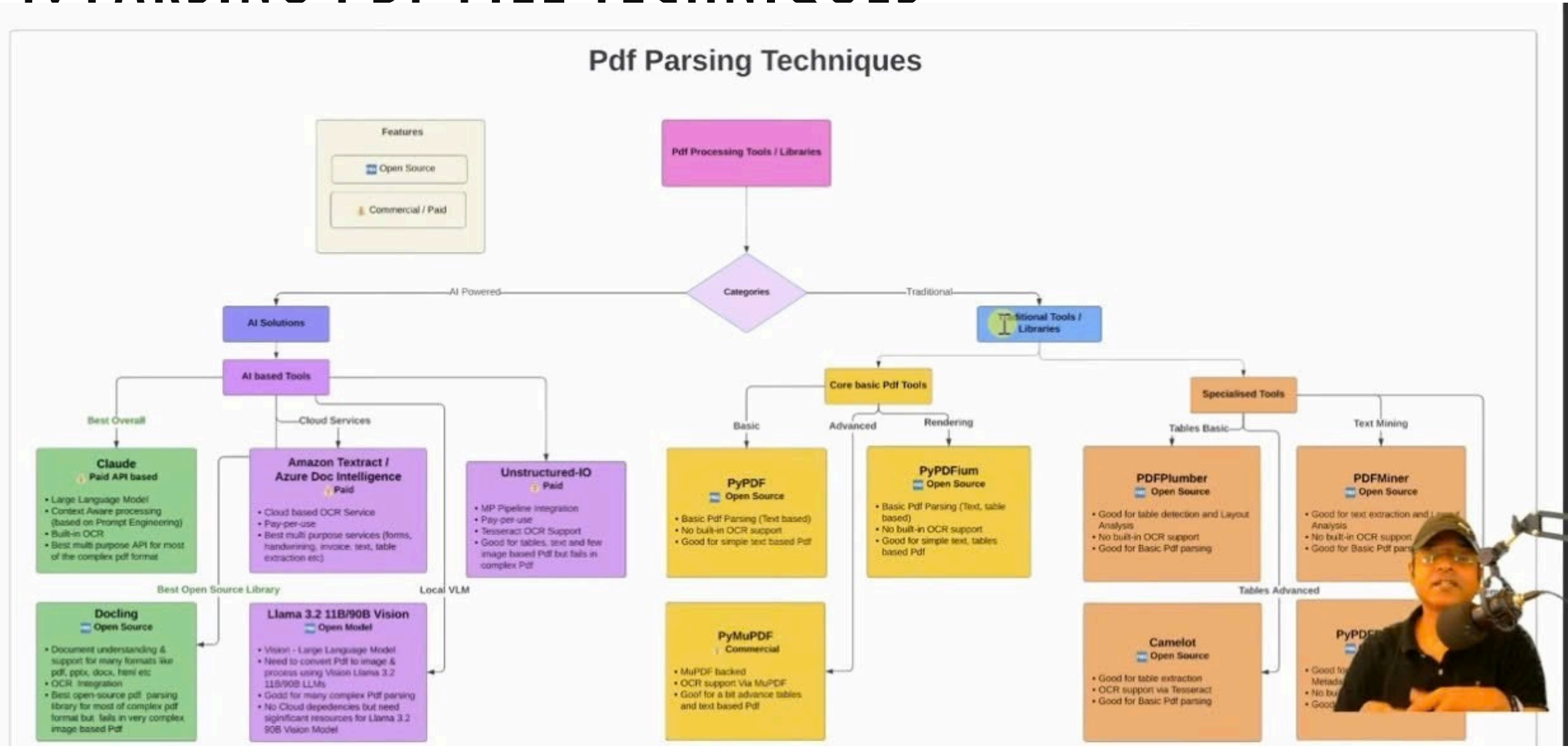
YenDT39

CONTENT

1. Parsing PDF File Techniques

2. Retrieving Information Techniques

1. PARSING PDF FILE TECHNIQUES



IBM DOCLING PARSING (OPEN SOURCE)

Open-source Python library by IBM for document processing

Parses diverse formats:

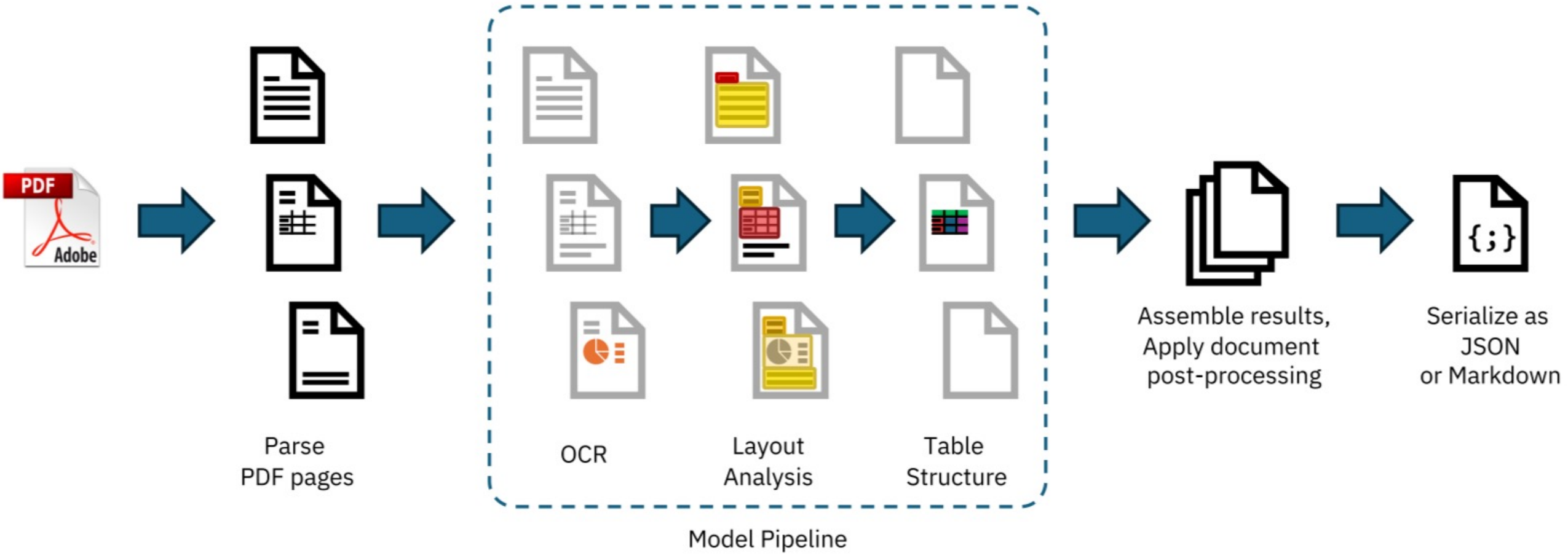
- Documents: PDF, DOCX, PPTX, XLSX, HTML
- Images: PNG, JPEG, TIFF
- Audio: WAV, MP3

Designed for generative AI workflows (e.g., RAG, chatbots)

Key benefits:

- Advanced parsing (layouts, tables, formulas)
- Local execution for secure data processing
- Seamless AI framework integrations

IBM DOCLING PARSING PIPELINE



MODELS USED IN DOCLING

- OCR Model
  - EasyOCR
  - Capturing small print detail in decent quality
- Layout Analysis Model
  - An object-detector which predicts the bounding-boxes and classes of various elements on the image of a given page. Its architecture is derived from RT-DETR [16] and re-trained on DocLayNet [13]
  - Group them into meaningful and complete units such as paragraphs, section titles, list items, captions, figures or tables
- Table Structure Recognition
  - The TableFormer model [12], first published in 2022 and since refined with a custom structure token language [9], is a vision-transformer model for table structure recovery
  - It can predict the logical row and column structure of a given table based on an input image, and determine which table cells belong to column headers, row headers or the table body.

Assembly -> Markdown representation

- Assembles all prediction results produced on each page into a well-defined datatype that encapsulates a converted document,
- Several algorithms augment features are applied, such as detection of the document language, correcting the reading order, matching figures with captions and labelling metadata such as title, authors and references

EXAMPLE - CONVERTING PDF TO MARKDOWN CODE

Converts PDF text, tables, and images to Markdown Ideal for feeding into AI pipelines

Consolidated Statement of Financial Position
Year ended 30 June 2019

J$'000	2018	2019
Assets		
Current Assets		
Inventories	2,199,273	3,225,686
Receivables and prepayments	2,302,693	2,585,519
Investment securities	269,530	130,385
Cash and short-term deposits	3,968,075	3,974,545
	8,739,571	**9,916,135**
Non-Current Assets		
Property, plant and equipment	6,775,727	6,724,278
Investment in associates	-	593,961
Loans receivable	-	165,545
Investment securities	215,760	379,060
	6,991,487	**7,862,844**
Total Assets	**15,731,058**	**17,778,979**
Liabilities		
Current Liabilities		
Trade and other payables	3,873,904	3,336,064
Short-term borrowings	376,686	485,724
Taxation payable	362,940	444,969
	4,613,530	**4,266,757**
Non-Current Liabilities		
Deferred tax liabilities	257,430	213,511
Borrowings	2,169,937	2,213,130
	2,427,367	**2,426,641**
Total Liabilities	**7,040,897**	**6,693,398**
Equity		
Capital and reserves attributable to the company's equity holders		
Share capital	1,192,647	1,192,647
Capital reserve	119,946	130,832
Translation reserve	30,086	29,048
Retained earnings	7,347,482	9,733,054
Total Equity	**8,690,161**	**11,085,581**
Total Liabilities and Equity	**15,731,058**	**17,778,979**

See accompanying notes.

Fetching 9 files: 0%| | 0/9 [00:00<?, ?it/s]

Consolulaler Slolemenl oi Flnancini Posiion Yedi einled J0 une 2019

J$'0oo	2018	2012
Assels		
Currenl Assels		
Invenlorles	2,199,273	3,225,686
Receivables and prepayments	2,902,693	2,585.519
Invesiment securiles	269,530	130,385
cash and sh@rt-lø M deposlls	3,968,075	3,974,545
	8,739,571	9,916,135
Non-Current Assets		
Property plant and equipment	6,775,727	6,724 278
InvesimentIn assocates		693,961
Loans recelvable		165,545
Invesient securlies	215.760	379,060
	6,991,487	7,862,844
Total Assels	15,731,050	17,778,979
Llabilives		
Curren[Liabilities		
Trade and other payables	3,873,904	3,336,064
Shon-Ierm borrowings	376,686	485,724
Taxabon payable	362,940	444,969
	4,617,5J0	4,266,757
Non-Current Liabilities		
Delerted tax liabilllies	257,430	213.511
Borrowings	2,169,937	2,213.130
	2,427,367	2,426,641
Total Liabililles	7,040,897	6,693,J98
Equity		
Capital and reserves attributabte to		
the company's equity holders		
Share capltal	1,192 647	1,192,647
Capital reserve	119,946	130 832
Translation reserve	30,086	29,048
Retained earnings	7,347,482	9,733.054
Total Equity	8,690,161	11,085.581
Total Liabilities and Equity	15.731.058	17.778,979

See accompanying notes

PwC Basic Understanding of a Company's Financials

Table 0

STRUCTURE DU MARCHE D'ASSURANCE & SON ACTIVITE

(M.D)

SOCIETES D'ASSURANCE RESIDENTES	FORME JURIDIQUE	SPECIALITE	PRIMES NETTES 2021	PRIMES NETTES 2022	Tx d'évolution 2022/2021
SOCIETES D'ASSURANCE DIRECTE					
STAR	SOCIETE ANONYME S.A	MULTI - BRANCHES	368,2	386,3	4,9%
COMAR	SOCIETE ANONYME S.A	MULTI - BRANCHES	233,3	252,8	8,4%
ASTREE	SOCIETE ANONYME S.A	MULTI - BRANCHES	187,2	236,0	26,1%
GAT	SOCIETE ANONYME S.A	MULTI - BRANCHES	218,4	235,0	7,6%
MAGHREBIA	SOCIETE ANONYME S.A	MULTI - BRANCHES	202,5	226,1	11,7%
ASSURANCES BIAT	SOCIETE ANONYME S.A	MULTI - BRANCHES	171,4	206,3	20,4%
AMI	SOCIETE ANONYME S.A	MULTI - BRANCHES	142,9	189,8	32,8%
BH ASSURANCE	SOCIETE ANONYME S.A	MULTI - BRANCHES	147,5	161,6	9,6%
LLOYD TUNISIEN	SOCIETE ANONYME S.A	MULTI - BRANCHES	144,4	159,3	10,3%
CARTE	SOCIETE ANONYME S.A	MULTI - BRANCHES	132,0	146,2	10,8%
MAE	SOCIETE MUTUELLE	MULTI - BRANCHES	151,9	164,7	8,4%
CTAMA	SOCIETE MUTUELLE	MULTI - BRANCHES	144,9	164,0	13,2%
ATTIJARI ASSURANCE	SOCIETE ANONYME S.A	VIE & CAPITALISATION	133,1	112,2	-15,7%
MAGHREBIA VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION	91,8	103,3	12,5%
CARTE VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION	66,1	100,6	52,2%
ASSURANCES HAYETT	SOCIETE ANONYME S.A	VIE & CAPITALISATION	80,6	83,1	3,1%
GAT VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION	36,9	53,9	46,1%
LLOYD VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION	6,4	8,4	31,3%
ASSURANCES UIB (1)	SOCIETE ANONYME S.A	VIE & CAPITALISATION	-	-	-
COTUNACE	SOCIETE ANONYME S.A	EXPORTATIONS & CREDITS	14,0	13,8	-1,4%
ZITOUNA TAKAFUL	SOCIETE ANONYME S.A	MULTI - BRANCHES	81,3	95,8	17,8%
EL AMANA TAKAFUL	SOCIETE ANONYME S.A	MULTI - BRANCHES	43,6	47,2	8,3%
ATTAKAFULIA	SOCIETE ANONYME S.A	MULTI - BRANCHES	34,8	38,6	10,9%
TOTAL 1			**2833,2**	**3185,0**	**12,4%**
SOCIETES DE REASSURANCE					
TUNIS - RE	SOCIETE ANONYME S.A	REASSURANCE	163,2	195,3	19,7%
TOTAL 2			**2996,4**	**3380,3**	**12,8%**

(1) : En 2020, l'agrément définitif a été octroyé à la nouvelle compagnie spécialisée en assurance vie et capitalisation «Assurances UIB» dont 'activité n'a pas encore démarré jusqu'à la fin de 2022

STRUCTURE DU MARCHE D'ASSURANCE & SON ACTIVITE

SOCIETES D'ASSURANCE RESIDENTES	FORME JURIDIQUE	SPECIALITE
PRIMES NETTES 2021 2022	PRIMES NETTES 2021 2022	(M.D) Tx d'évolution 2022/2021
SOCIETES D'ASSURANCE DIRECTE	SOCIETES D'ASSURANCE DIRECTE	SOCIETES D'ASSURANCE DIRECTE
SOCIETES D'ASSURANCE DIRECTE	SOCIETES D'ASSURANCE DIRECTE	SOCIETES D'ASSURANCE DIRECTE
STAR	SOCIETE ANONYME S.A	MULTI BRANCHES
368,2	386,3	4,9%
COMAR	SOCIETE ANONYME S.A	MULTI BRANCHES
233,3	252,8	8,4%
ASTREE	SOCIETE ANONYME S.A	MULTI BRANCHES
187,2	236,0	26,1 %
GAT	SOCIETE ANONYME S.A	MULTI BRANCHES
218,4	235,0	7,6%
MAGHREBIA	SOCIETE ANONYME S.A	MULTI BRANCHES
202,5	226,1	11,7%
ASSURANCES BIAT	SOCIETE ANONYME S.A	MULTI BRANCHES
171,4	206,3	20,4%
AMI	SOCIETE ANONYME S.A	MULTI BRANCHES
142,9	189,8	32,8%
BH ASSURANCE	SOCIETE ANONYME S.A	MULTI BRANCHES
147,5	161,6	9,6%
LLOYD TUNISIEN	SOCIETE ANONYME S.A	MULTI BRANCHES
144,4	159,3	10,3%
CARTE	SOCIETE ANONYME S.A	MULTI BRANCHES
132,0	146,2	10,8%
MAE	SOCIETE MUTUELLE	MULTI BRANCHES
151,9	164,7	8,4%
CTAMA	SOCIETE MUTUELLE	MULTI BRANCHES
144,9	164,0	13,2%
ATTIJARI ASSURANCE	SOCIETE ANONYME S.A	VIE & CAPITALISATION
133,1	112,2	415,7%
MAGHREBIA VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION
91,8	103,3	12,5%
CARTE VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION
66,1	100,6	52,2%
ASSURANCES HAYETT	SOCIETE ANONYME S.A	VIE & CAPITALISATION
80,6	83,1	3,1 %
GAT VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION
36,9	53,9	46,1 %
LLOYD VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION
6,4	8,4	31,3%
ASSURANCES UIB	SOCIETE ANONYME S.A	VIE & CAPITALISATION
COTUNACE	SOCIETE ANONYME S.A	EXPORTATIONS & CREDITS

```
df = pd.read_csv("/kaggle/working/other_table_big-table-1.csv")
df = df.loc[:, ~df.columns.str.contains('^Unnamed')]
df
```

	SOCIETES D'ASSURANCE RESIDENTES	FORME JURIDIQUE	SPECIALITE	PRIMES NETTES 2021 2022	PRIMES NETTES 2021 2022.1	(M.D) Tx d'évolution 2022/2021
0	SOCIETES D'ASSURANCE DIRECTE	SOCIETES D'ASSURANCE DIRECTE	SOCIETES D'ASSURANCE DIRECTE	SOCIETES D'ASSURANCE DIRECTE	SOCIETES D'ASSURANCE DIRECTE	SOCIETES D'ASSURANCE DIRECTE
1	STAR	SOCIETE ANONYME S.A	MULTI BRANCHES	368,2	386,3	4,9%
2	COMAR	SOCIETE ANONYME S.A	MULTI BRANCHES	233,3	252,8	8,4%
3	ASTREE	SOCIETE ANONYME S.A	MULTI BRANCHES	187,2	236,0	26,1 %
4	GAT	SOCIETE ANONYME S.A	MULTI BRANCHES	218,4	235,0	7,6%
5	MAGHREBIA	SOCIETE ANONYME S.A	MULTI BRANCHES	202,5	226,1	11,7%
6	ASSURANCES BIAT	SOCIETE ANONYME S.A	MULTI BRANCHES	171,4	206,3	20,4%
7	AMI	SOCIETE ANONYME S.A	MULTI BRANCHES	142,9	189,8	32,8%
8	BH ASSURANCE	SOCIETE ANONYME S.A	MULTI BRANCHES	147,5	161,6	9,6%
9	LLOYD TUNISIEN	SOCIETE ANONYME S.A	MULTI BRANCHES	144,4	159,3	10,3%
10	CARTE	SOCIETE ANONYME S.A	MULTI BRANCHES	132,0	146,2	10,8%
11	MAE	SOCIETE MUTUELLE	MULTI BRANCHES	151,9	164,7	8,4%
12	CTAMA	SOCIETE MUTUELLE	MULTI BRANCHES	144,9	164,0	13,2%
13	ATTIJARI ASSURANCE	SOCIETE ANONYME S.A	VIE & CAPITALISATION	133,1	112,2	415,7%
14	MAGHREBIA VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION	91,8	103,3	12,5%
15	CARTE VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION	66,1	100,6	52,2%
16	ASSURANCES HAYETT	SOCIETE ANONYME S.A	VIE & CAPITALISATION	80,6	83,1	3,1 %
17	GAT VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION	36,9	53,9	46,1 %
18	LLOYD VIE	SOCIETE ANONYME S.A	VIE & CAPITALISATION	6,4	8,4	31,3%
19	ASSURANCES UIB	SOCIETE ANONYME S.A	VIE & CAPITALISATION	NaN	NaN	NaN
20	COTUNACE	SOCIETE ANONYME S.A	EXPORTATIONS & CREDITS	14,0	13,8	-1,4%
21	ZITOUNA TAKAFUL	SOCIETE ANONYME S.A	MULTI BRANCHES	81,3	95,8	17,8%
22	ELAMANA TAKAFUL	SOCIETE ANONYME S.A	MULTI BRANCHES	43,6	47,2	8,3%
23	ATTAKAFULIA	SOCIETE ANONYME S.A	MULTI BRANCHES	34,8	38,6	10,9%
24	TOTAL 1	TOTAL 1	TOTAL 1	2833,2	3185,0	12,4%
25	SOCIETES DE REASSURANCE	SOCIETES DE REASSURANCE	SOCIETES DE REASSURANCE	SOCIETES DE REASSURANCE	SOCIETES DE REASSURANCE	SOCIETES DE REASSURANCE
26	TUNIS RE	SOCIETE ANONYME S.A	REASSURANCE	163,2	195,3	19,7%
27	TOTAL 2	TOTAL 2	TOTAL 2	2996,4	3380,3	12,8%

COSTCO WHOLESALE CORPORATION
CONSOLIDATED STATEMENTS OF EQUITY
(amounts in millions)

	Common Stock		Additional Paid-in Capital	Accumulated Other Comprehensive Income (Loss)	Retained Earnings	Total Costco Stockholders' Equity	Noncontrolling Interests	Total Equity
	Shares (000's)	Amount						
BALANCE AT AUGUST 30, 2020	441,255	$ 4	$ 6,698	$ (1,297)	$ 12,879	$ 18,284	$ 421	$ 18,705
Net income	—	—	—	—	5,007	5,007	72	5,079
Foreign-currency translation adjustment and other, net	—	—	—	160	—	160	21	181
Stock-based compensation	—	—	668	—	—	668	—	668
Release of vested restricted stock units (RSUs), including tax effects	1,928	—	(312)	—	—	(312)	—	(312)
Repurchases of common stock	(1,358)	—	(23)	—	(472)	(495)	—	(495)
Cash dividends declared	—	—	—	—	(5,748)	(5,748)	—	(5,748)
BALANCE AT AUGUST 29, 2021	441,825	4	7,031	(1,137)	11,666	17,564	514	18,078
Net income	—	—	—	—	5,844	5,844	71	5,915
Foreign-currency translation adjustment and other, net	—	—	—	(686)	—	(686)	(35)	(721)
Stock-based compensation	—	—	728	—	—	728	—	728
Release of vested RSUs, including tax effects	1,702	—	(363)	—	—	(363)	—	(363)
Dividend to noncontrolling interest	—	—	—	—	—	—	(208)	(208)
Acquisition of noncontrolling interest	—	—	(499)	(6)	—	(505)	(337)	(842)
Repurchases of common stock	(863)	—	(15)	—	(427)	(442)	—	(442)
Cash dividends declared and other	—	(2)	2	—	(1,498)	(1,498)	—	(1,498)
BALANCE AT AUGUST 28, 2022	442,664	2	6,884	(1,829)	15,585	20,642	5	20,647
Net income	—	—	—	—	6,292	6,292	—	6,292
Foreign-currency translation adjustment and other, net	—	—	—	24	—	24	—	24
Stock-based compensation	—	—	778	—	—	778	—	778
Release of vested RSUs, including tax effects	1,470	—	(303)	—	—	(303)	—	(303)
Repurchases of common stock	(1,341)	—	(24)	—	(653)	(677)	—	(677)
Cash dividends declared and other	—	—	5	—	(1,703)	(1,698)	(5)	(1,703)
BALANCE AT SEPTEMBER 3, 2023	442,793	$ 2	$ 7,340	$ (1,805)	$ 19,521	$ 25,058	$ —	$ 25,058

```
## COSTCO WHOLESALE CORPORATION

## CONSOLIDATED STATEMENTS OF EQUITY

(amounts in millions)

|                                                                      | Common Stock   | Com
mon Stock   | Additional   | Accumulated Other   |          | Total Costco         |
|          |
|----------------------------------------------------------------------|---------------|----
------------|-------------|--------------------|----------|---------------------|---------
------|---------|
|                                                                      | Shares         |
| Paid-in      | Comprehensive       | Retained |                     | Noncontrolling | Tota
l    |
|                                                                      | (oo0's)        | Amo
unt          | Capital      | Income (Loss)       | Earnings | Stockholders' Equity | Interests
| Equity   |
| BALANCE AT AUGUST 30, 2020                                           | 441,255        | $ 4
| $ 6,698      | $ (1,297)           | $ 12,879 | $ 18,284             | $ 421          | $ 1
8,705 |
| Net income                                                           |                |
|              |                     | 5,007    | 5,007                | 72             | 5,07
9    |
| Foreign-currency translation adjustment and other; net               |                |
|              | 160                 |          | 160                  | 21             | 181
|
| Stock-based compensation                                             |                |
| 668          |                     |          | 668                  |                | 668
|
| Release of vested restricted stock units (RSUs) including tax effects | 1,928         |
| (312)        |                     |          | (312)                |                | (31
2)    |
| Repurchases of common stock                                          | (1,358)        |
| (23)         |                     | (472)    | (495)                |                | (49
5)    |
| Cash dividends declared                                              |                |
|              |                     | (5,748)  | (5,748)              |                | (5,7
48)  |
| BALANCE AT AUGUST 29, 2021                                           | 441,825        |
| 7,031        | (1,137)             | 11,666   | 17,564               | 514            | 18,0
78   |
| Net income                                                           |                |
|              |                     | 5,844    | 5,844                | 71             | 5,91
5    |
| Foreign-currency translation adjustment and other; net               |                |
|              | (686)               |          | (686)                | (35)           | (72
1)    |
| Stock-based compensation                                             |                |
| 728          |                     |          | 728                  |                | 728
|
| Release of vested RSUs, including tax effects                        | 1,702          |
| (363)        |                     |          | (363)                |                | (36
```

```python
df = pd.read_csv("/kaggle/working/table_3-table-1.csv")
df = df.loc[:, ~df.columns.str.contains('^Unnamed: 0')]
df
```

	Unnamed: 1	Common Stock.Shares.(oo0's)	Common Stock..Amount	Additional.Paid-in.Capital	Accumulated Other.Comprehensive.Income (Loss)	Retained.Earnings	Total Costco..Stockholders Equity
0	BALANCE AT AUGUST 30, 2020	441,255	$ 4	$ 6,698	$ (1,297)	$ 12,879	$ 18,284
1	Net income	NaN	NaN	NaN	NaN	5,007	5,007
2	Foreign-currency translation adjustment and ot...	NaN	NaN	NaN	160	NaN	160
3	Stock-based compensation	NaN	NaN	668	NaN	NaN	668
4	Release of vested restricted stock units (RSUs...	1,928	NaN	(312)	NaN	NaN	(312)
5	Repurchases of common stock	(1,358)	NaN	(23)	NaN	(472)	(495)
6	Cash dividends declared	NaN	NaN	NaN	NaN	(5,748)	(5,748)
7	BALANCE AT AUGUST 29, 2021	441,825	NaN	7,031	(1,137)	11,666	17,564
8	Net income	NaN	NaN	NaN	NaN	5,844	5,844
9	Foreign-currency translation adjustment and ot...	NaN	NaN	NaN	(686)	NaN	(686)
10	Stock-based compensation	NaN	NaN	728	NaN	NaN	728
11	Release of vested RSUs, including tax effects	1,702	NaN	(363)	NaN	NaN	(363)
12	Dividend to noncontrolling interest	NaN	NaN	NaN	NaN	NaN	NaN
13	Acquisition of noncontrolling interest	NaN	NaN	(499)	(6)	NaN	(505)
14	Repurchases of common stock	(863)	NaN	(15)	NaN	(427)	(442)
15	Cash dividends declared and other	NaN	'2)	2	NaN	1,498)	(1,498)
16	BALANCE AT AUGUST 28, 2022	442,664	2	6,884	(1,829)	15,585	20,642
17	Net income	NaN	NaN	NaN	NaN	6,292	6,292
18	Foreign-currency translation adjustment and ot...	NaN	NaN	NaN	24	NaN	24
19	Stock-based compensation	NaN	NaN	778	NaN	NaN	778
20	Release of vested RSUs, including tax effects	1,470	NaN	(303)	NaN	NaN	(303)

The following table sets forth, for the periods indicated, certain items in our Consolidated Statements of Income expressed as a percentage of revenue.

	Year Ended	
	Jan 28, 2024	Jan 29, 2023
Revenue	100.0 %	100.0 %
Cost of revenue	27.3	43.1
Gross profit	72.7	56.9
Operating expenses		
Research and development	14.2	27.2
Sales, general and administrative	4.4	9.1
Acquisition termination cost	—	5.0
Total operating expenses	18.6	41.3
Operating income	54.1	15.6
Interest income	1.4	1.0
Interest expense	(0.4)	(1.0)
Other, net	0.4	(0.1)
Other income (expense), net	1.4	(0.1)
Income before income tax	55.5	15.5
Income tax expense (benefit)	6.6	(0.7)
Net income	48.9 %	16.2 %

Reportable Segments

Revenue by Reportable Segments

	Year Ended			
	Jan 28, 2024	Jan 29, 2023	$ Change	% Change
	($ in millions)			
Compute & Networking	$ 47,405	$ 15,068	$ 32,337	215 %
Graphics	13,517	11,906	1,611	14 %
Total	$ 60,922	$ 26,974	$ 33,948	126 %

Operating Income by Reportable Segments

	Year Ended			
	Jan 28, 2024	Jan 29, 2023	$ Change	% Change
	($ in millions)			
Compute & Networking	$ 32,016	$ 5,083	$ 26,933	530 %
Graphics	5,846	4,552	1,294	28 %
All Other	(4,890)	(5,411)	521	(10)%
Total	$ 32,972	$ 4,224	$ 28,748	681 %

Compute & Networking revenue – The year-on-year increase was due to higher Data Center revenue. Compute grew 266% due to higher shipments of the NVIDIA Hopper GPU computing platform for the training and inference of LLMs, recommendation engines and generative AI applications. Networking was up 133% due to higher shipments of InfiniBand.

Graphics revenue – The year-on-year increase was led by growth in Gaming of 15% driven by higher sell-in to partners following the normalization of channel inventory levels.

Reportable segment operating income – The year-on-year increase in Compute & Networking and Graphics operating income was driven by higher revenue.


```
The following table sets forth, for the periods indicated, certain items in our Consolidated S
tatements of Income expressed as a percentage of revenue.

|                                  | Year Ended   | Year Ended   |
|----------------------------------|--------------|--------------|
|                                  | Jan 28, 2024 | Jan 29, 2023 |
| Revenue                          | 100.0 %      | 100.0 %      |
| Cost of revenue                  | 27.3         | 43.1         |
| Gross profit                     | 72.7         | 56.9         |
| Operating expenses               |              |              |
| Research and development         | 14.2         | 27.2         |
| Sales, general and administrative | 4.4          | 9.1          |
| Acquisition termination cost     | -            | 5.0          |
| Total operating expenses         | 18.6         | 41.3         |
| Operating income                 | 54.1         | 15.6         |
| Interest income                  | 1.4          | 1.0          |
| Interest expense                 | (0.4)        | (1.0)        |
| Other, net                       | 0.4          | (0.1)        |
| Other income (expense), net      | 1.4          | (0.1)        |
| Income before income tax         | 55.5         | 15.5         |
| Income tax expense (benefit)     | 6.6          | (0.7)        |
| Net income                       | 48.9 %       | 16.2 %       |

## Reportable Segments

## Revenue by Reportable Segments

|                     | Year Ended      | Year Ended      | Year Ended      | Year Ended
|
|---------------------|-----------------|-----------------|-----------------|---------------
-|
|                     | Jan 28, 2024    | Jan 29, 2023    | $ Change        | % Change
|
|                     | ($ in millions) | ($ in millions) | ($ in millions) | ($ in millions)
|
| Compute & Networking | $  47,405       | $  15,068       | $  32,337       | 215 %
|
| Graphics            | 13,517          | 11,906          | 1,611           | 14 %
|
| Total               | $  60,922       | $  26,974       | $  33,948       | 126 %
|

## Operating Income by Reportable Segments

|                     | Year Ended      | Year Ended      | Year Ended      | Year Ended
|
|---------------------|-----------------|-----------------|-----------------|---------------
-|
|                     | Jan 28, 2024    | Jan 29, 2023    | $ Change        | % Change
|
|                     | ($ in millions) | ($ in millions) | ($ in millions) | ($ in millions)
|
| Compute & Networking | $  32,016       | $  5,083        | $  26,933       | 530 %
|
| Graphics            | 5,846           | 4,552           | 1,294           | 28 %
|
| All Other           | (4,890)         | (5,411)         | 521             | (10)%
|
| Total               | $  32,972       | $  4,224        | $  28,748       | 681 %
|

Compute & Networking revenue - The year-on-year increase was due to higher Data Center revenu
e. Compute grew 266% due to higher shipments of the NVIDIA Hopper GPU computing platform for t
he training and inference of LLMs, recommendation engines and generative AI applications. Netw
```

```
]: df = pd.read_csv("/kaggle/working/NVIDIA_Report_Removed_Pages-1-16-9-table-1.csv")
   df = df.loc[:, ~df.columns.str.contains('^Unnamed: 0')]
   df
```

]:

	Unnamed: 1	Year Ended.Jan 28, 2024	Year Ended.Jan 29, 2023
0	Revenue	100.0 %	100.0 %
1	Cost of revenue	27.3	43.1
2	Gross profit	72.7	56.9
3	Operating expenses	NaN	NaN
4	Research and development	14.2	27.2
5	Sales, general and administrative	4.4	9.1
6	Acquisition termination cost	-	5.0
7	Total operating expenses	18.6	41.3
8	Operating income	54.1	15.6
9	Interest income	1.4	1.0
10	Interest expense	(0.4)	(1.0)
11	Other, net	0.4	(0.1)
12	Other income (expense), net	1.4	(0.1)
13	Income before income tax	55.5	15.5
14	Income tax expense (benefit)	6.6	(0.7)
15	Net income	48.9 %	16.2 %

```
]: df = pd.read_csv("/kaggle/working/NVIDIA_Report_Removed_Pages-1-16-9-table-2.csv")
   df = df.loc[:, ~df.columns.str.contains('^Unnamed: 0')]
   df
```

]:

	Unnamed: 1	Year Ended.Jan 28, 2024.($ in millions)	Year Ended.Jan 29, 2023.($ in millions)	Year Ended.Change. (in millions)	Year Ended.% Change.($ in millions)
0	Compute & Networking	$ 47,405	$ 15,068	$ 32,337	215 %
1	Graphics	13,517	11,906	1,611	14 %
2	Total	$ 60,922	$ 26,974	$ 33,948	126 %

```
]: df = pd.read_csv("/kaggle/working/NVIDIA_Report_Removed_Pages-1-16-9-table-3.csv")
   df = df.loc[:, ~df.columns.str.contains('^Unnamed: 0')]
   df
```

]:

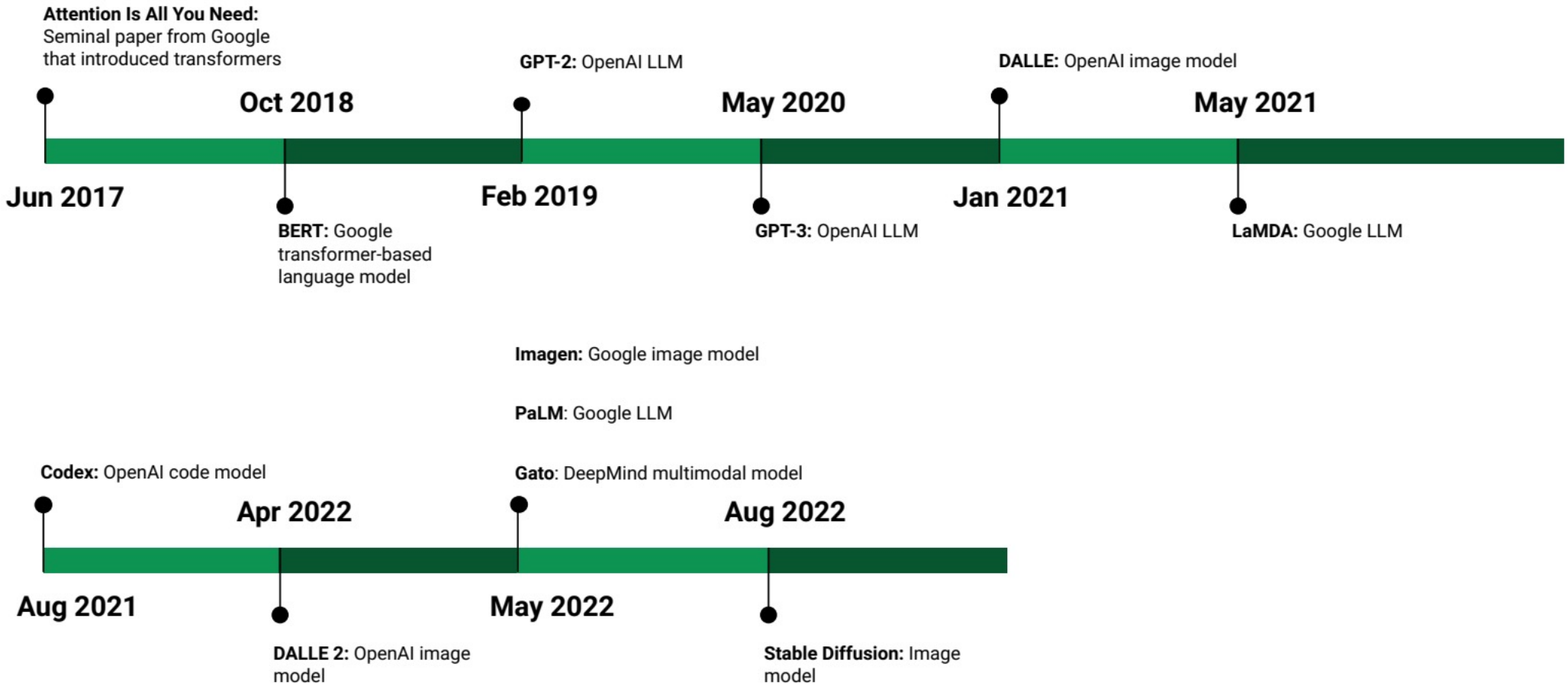
	Unnamed: 1	Year Ended.Jan 28, 2024.($ in millions)	Year Ended.Jan 29, 2023.($ in millions)	Year Ended.Change. (in millions)	Year Ended.% Change.($ in millions)
0	Compute & Networking	$ 32,016	$ 5,083	$ 26,933	530 %
1	Graphics	5,846	4,552	1,294	28 %
2	All Other	(4,890)	(5,411)	521	(10)%
3	Total	$ 32,972	$ 4,224	$ 28,748	681 %

2. RETRIEVING INFORMATION

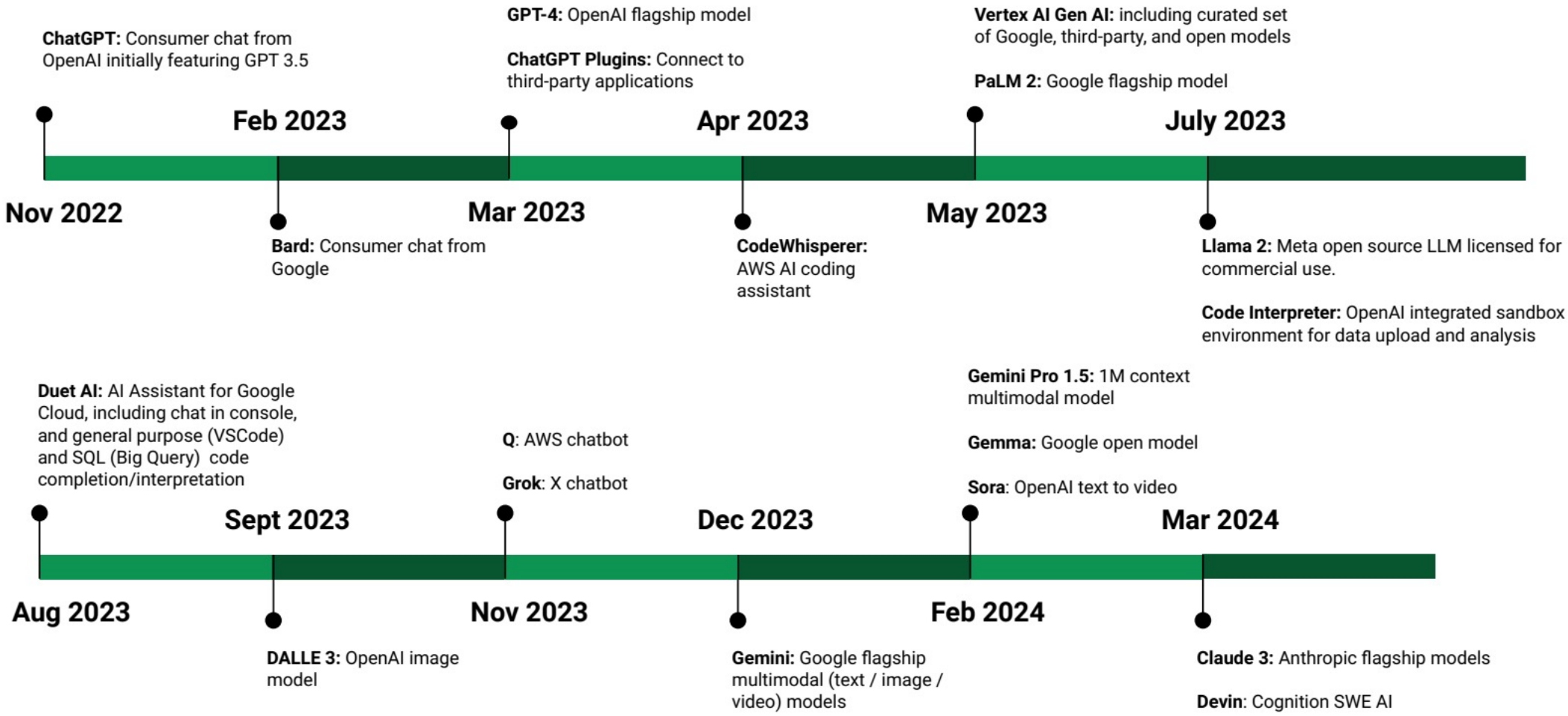
2.1. Major Generative AI Milestones

2.2. LLM and RAG for Retrieving Information

2.1 MAJOR GENERATIVE AI MILESTONES: PART 1



2.1 MAJOR GENERATIVE AI MILESTONES: PART 2

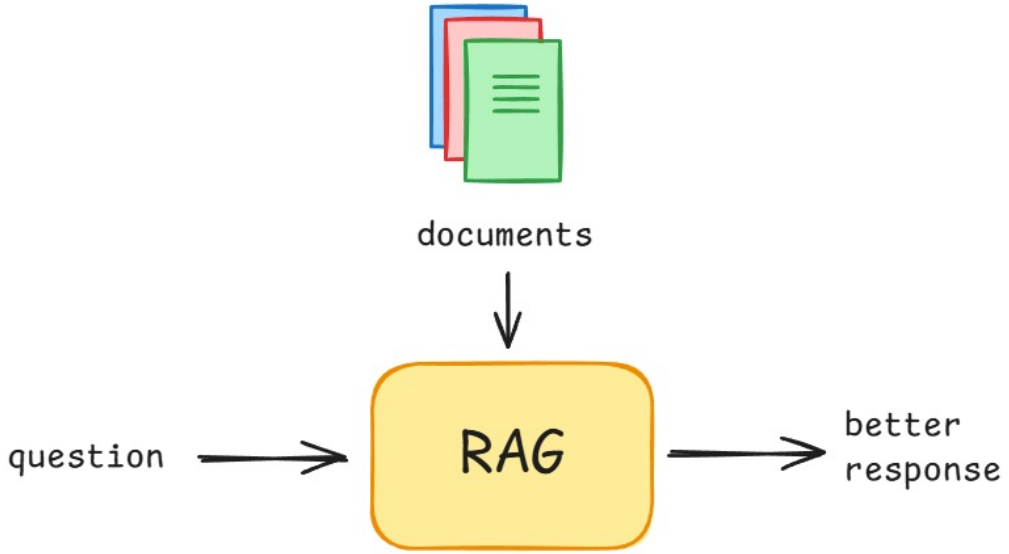


2.1 GENERATIVE AI LANDSCAPE BY VENDOR

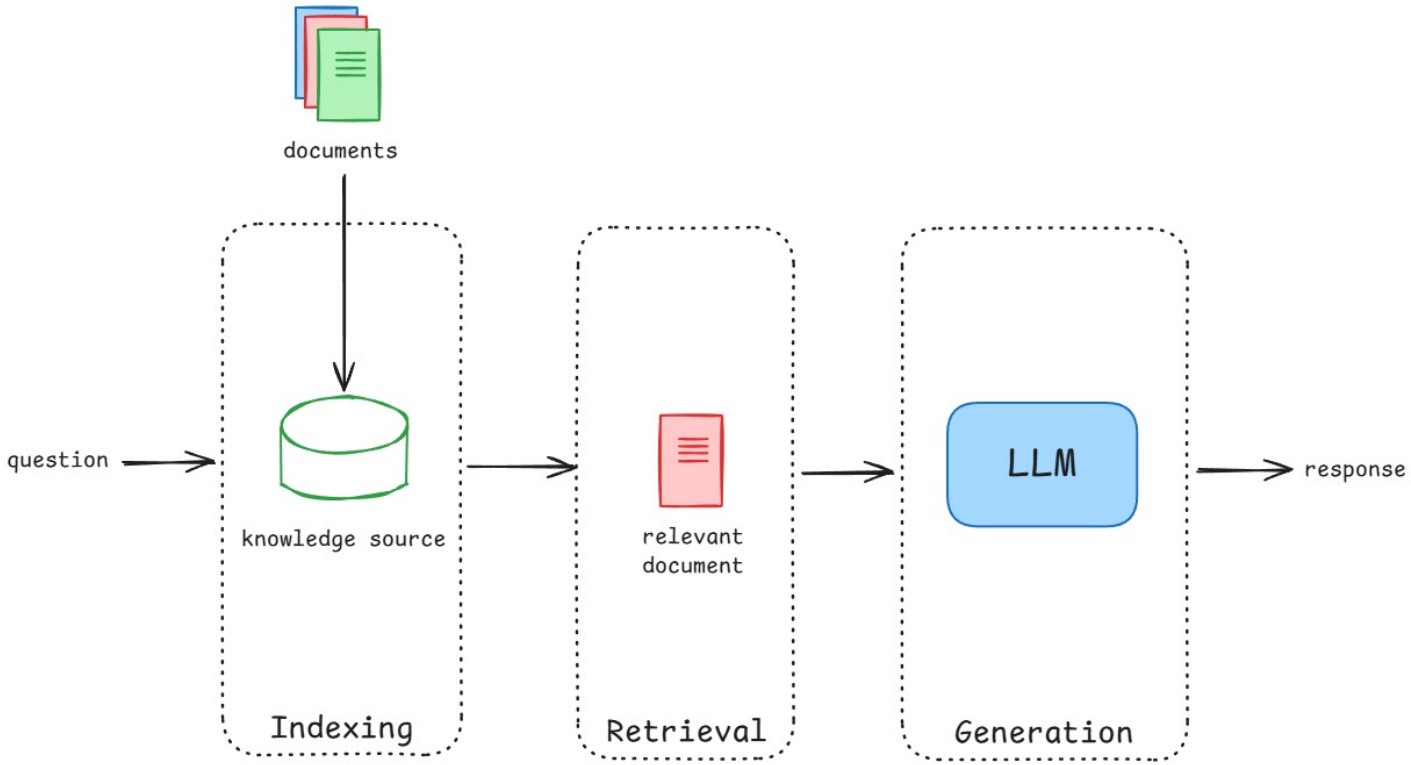
Vendor	Prod. Suite Assistance	Developer / Ops Assistant	Consumer Chat	Enterprise Gen AI	Dev / Hobbyist Gen AI	Open Foundation Models
Google	Gemini for Google Workspace	Duet AI for Google Cloud	Gemini	Vertex AI	Google AI for Developers	Gemma
Microsoft	CoPilot 365	Github Copilot	Bing Chat	Azure OpenAI		
OpenAI		ChatGPT	ChatGPT	ChatGPT Enterprise	ChatGPT	
AWS		• Q • CodeWhisperer		Bedrock / Titan		
Anthropic			Claude 3*	Claude 3*		
Meta						Llama 2*
Mistral						Mixtral 8x7B*

2.2 LLM AND RAG (RETRIEVAL AUGMENTED GENERATION)

Large Language Models (LLMs), like GPT, Gemini, or Llama, have revolutionized AI by enabling natural language understanding and generation at an unprecedented scale. Despite being incredibly powerful, LLMs have notable limitations.



HOW RAG WORKS



A RAG pipeline usually has three main components:

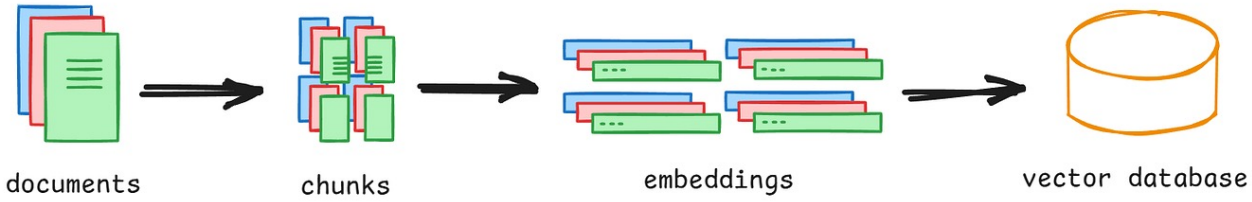
- A knowledge source (indexing),
- A retriever,
- A generative model.

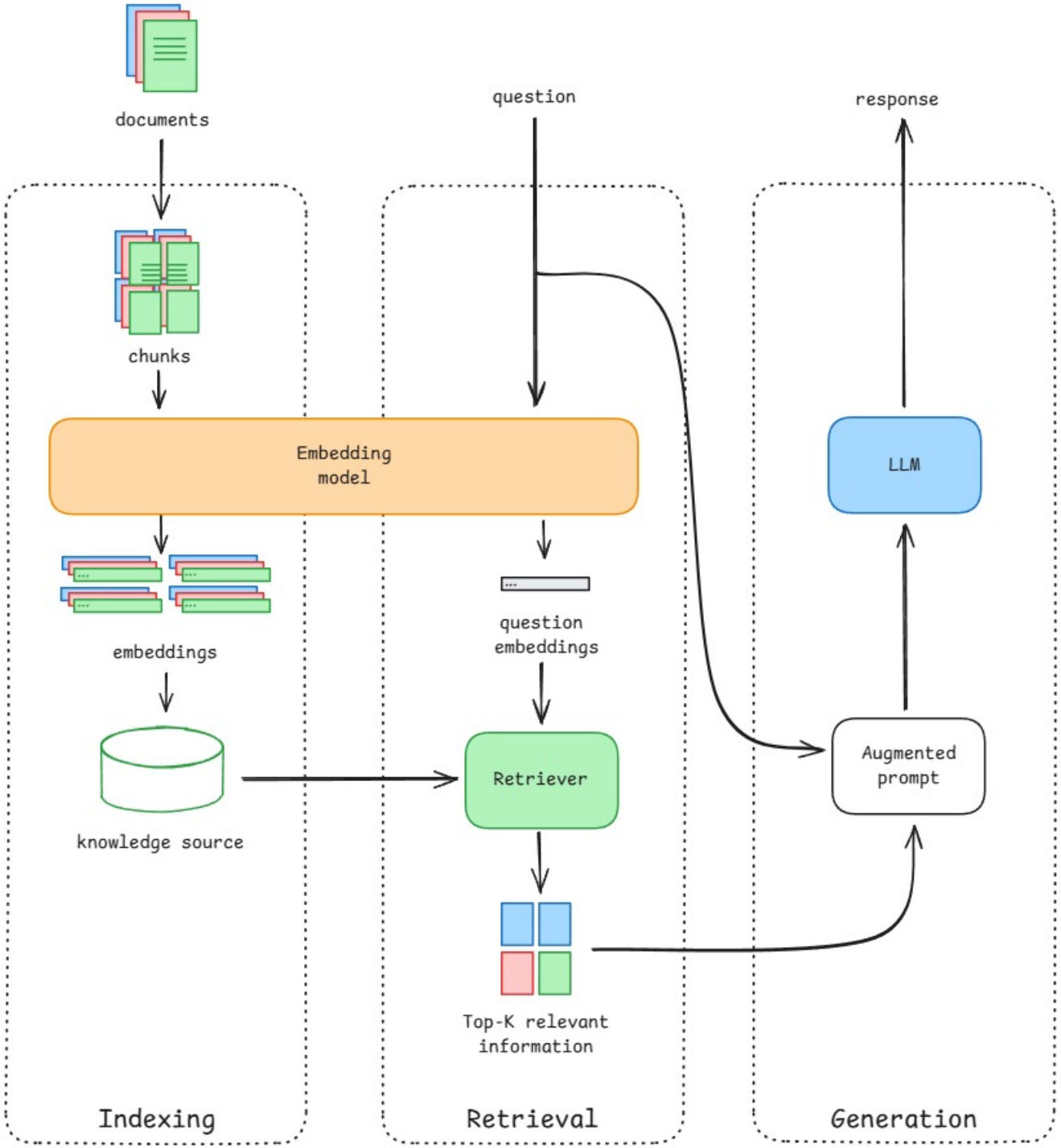
The knowledge source: each document is split into small pieces, called chunks, in order to break down information into more manageable pieces.

The chunks are then transformed into embeddings using an embedding model. These embeddings are either stored locally or in a vector database, ready for quick retrieval

A similarity search between two embeddings, distance between two embeddings.

Top-K retrieval is simply returning the top-K chunks of documents with the smallest distance to the query in vector space.





Indexing

- Data Parsing
- Data Chunking
- Data Embedding
- Indexing
- Storing

Retrieval

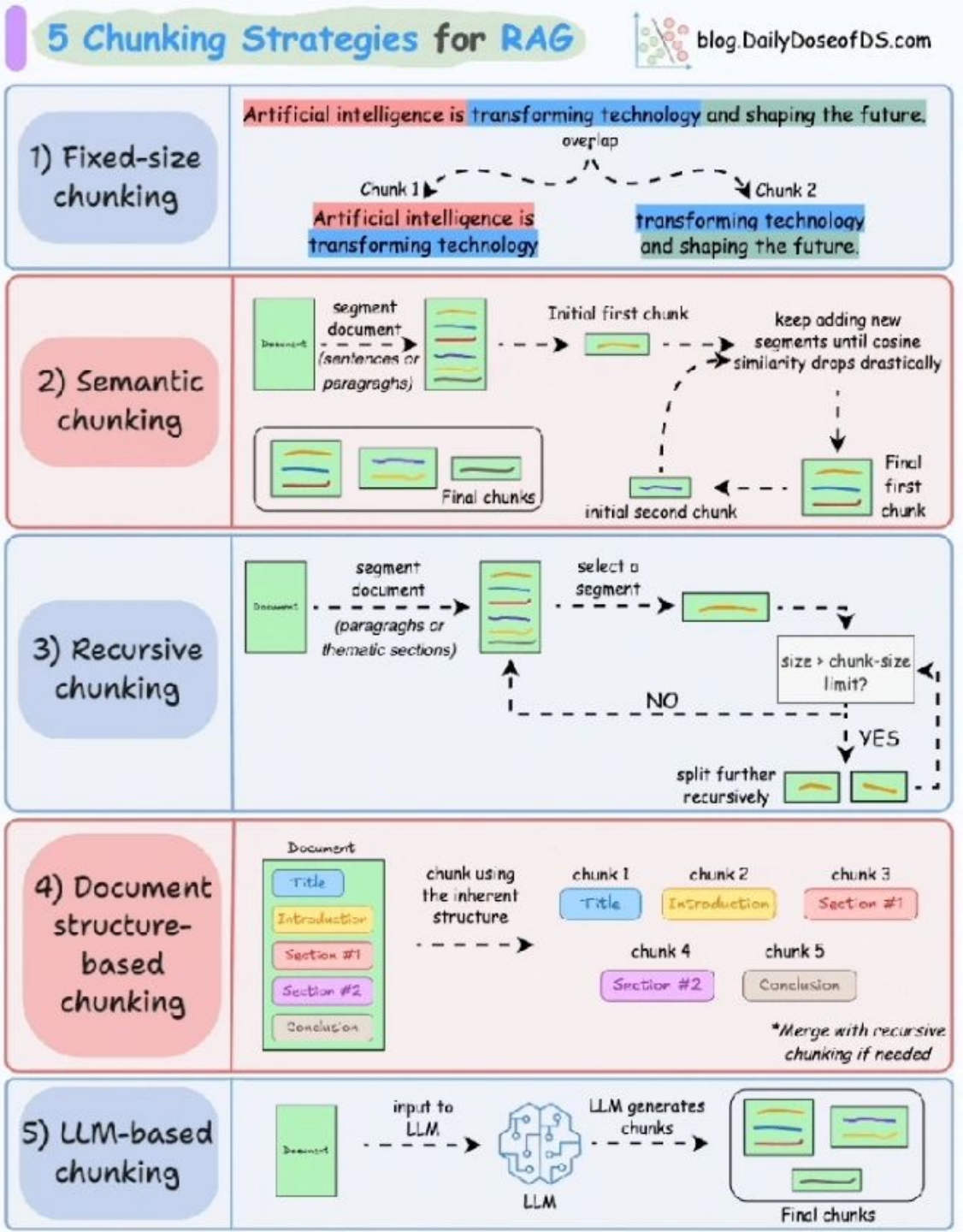
- Retriever
- Top K relevant information

Generation

- Top K relevant information are plugged (prompt) into LLM response synthesis

DATA CHUNKING

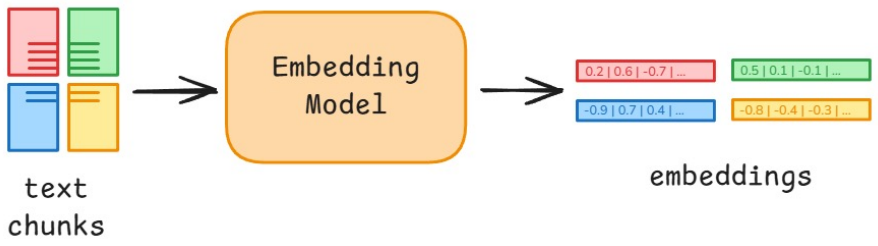
- **Fixed-size chunking:** breaks down the text into chunks of a predetermined size, typically measured in characters, words, or tokens (e.g., 1000 characters)
- **Semantic chunking:** Semantic chunking goes beyond structure and focuses on meaning. It analyzes the semantic relationships between parts of the text to determine where to split (similarity)
- **Recursive chunking:** structured approach. It starts by splitting the text using a primary separator, such as paragraphs. Paragraphs first, then sentences, and finally individual phrases.
- **Document-based chunking:** This method treats each logical section of a document — like headings, paragraphs, or subsections — as a chunk.
- **Agentic chunking:** This advanced technique uses LLMs to dynamically split text based on meaning, structure, and context. An LLM evaluates the document to identify the best points to divide the content, often isolating ideas or propositions into their own chunks.



DATA EMBEDDING & STORING

Embedding

- Embedding models don’t just convert text into vectors; they capture the context within the text, preserving its meaning.
- Use pretrained model or fine-tuning embedding models



Vector Database

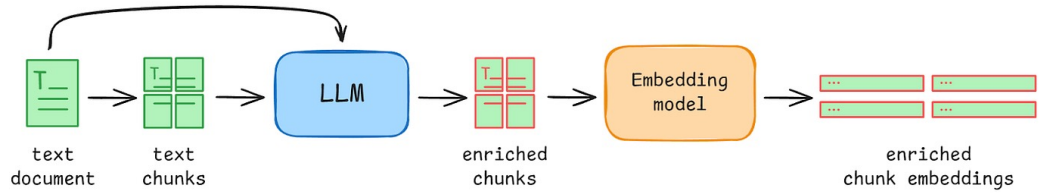
Using libraries like FAISS that specialize in similarity searches, Vector DBs (like Pinecone, Weaviate, and Milvus) extend their capabilities by offering a full-fledged data management system. They enable CRUD (Create, Read, Update, Delete) operations, real-time updates, metadata integration, high storage volumes, and robust security, making them ideal for production-grade applications.

Data Indexing

Contextual Retrieval:

Each chunk is paired with the full document and sent to the LLM, which generates a context-enriched version of the chunk.

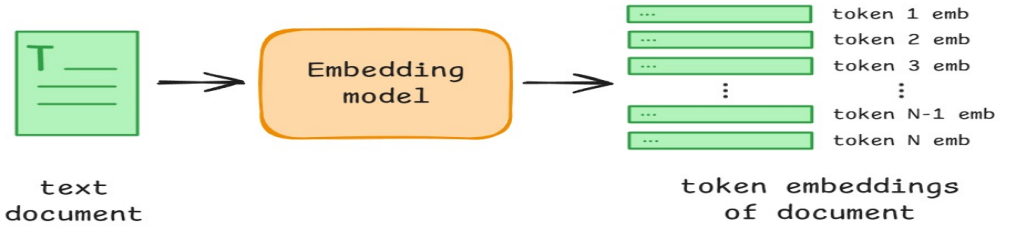
The enriched **chunks** are embedded into **vectors**, reflecting both the chunk's local content and relevant global context



Late Interaction:

Each document is embedded into **token-level** representations, with all token embeddings stored in a vector index.

Using MaxSim to compare tokens of query and document

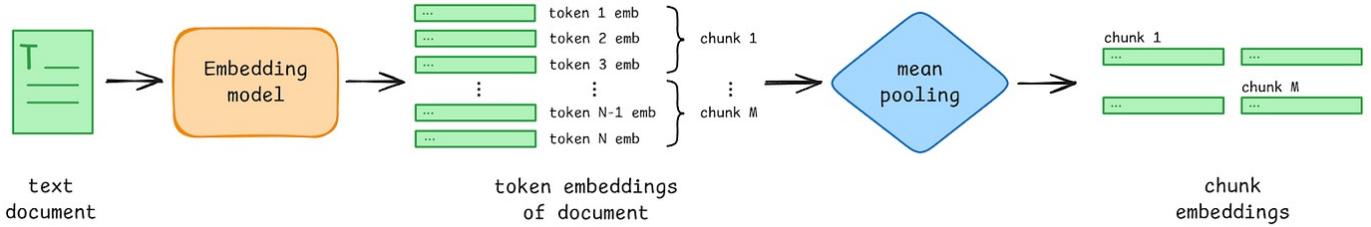


Late Chunking:

The entire document is embedded into token-level

Chunk boundaries (e.g., sentences or sections) are identified, and their positions are noted.

Chunks are then derived from the token embeddings using mean pooling, ensuring each chunk maintains the contextual richness of the full document.



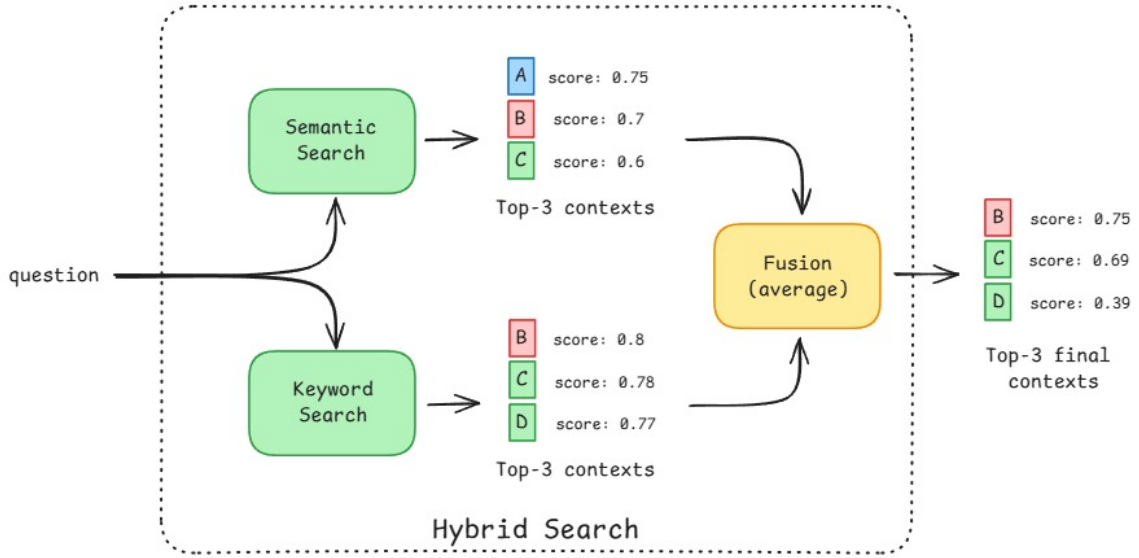
RETRIEVAL: SIMILARITY SEARCH

- Semantic search: Ensures context and intent are captured: Measure similarity between these vectors to find semantically similar matches (Not exact terms)

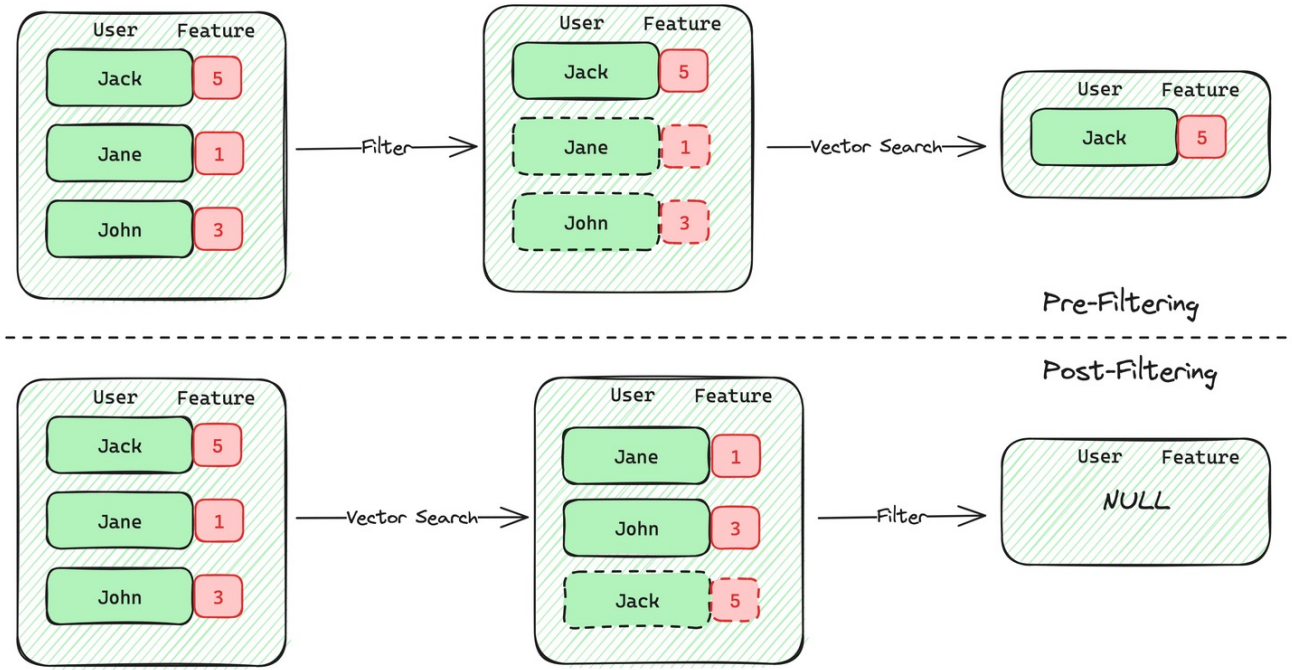
Keyword search: Guarantees critical terms are not overlooked (using like TF-IDF or BM25 rank documents based on the importance and frequency of terms)

Hybrid search: Results form both methods are merged and weighted to balance their contributions

Final Score = (Keyword Score × α) + (Semantic Score × (1−α))



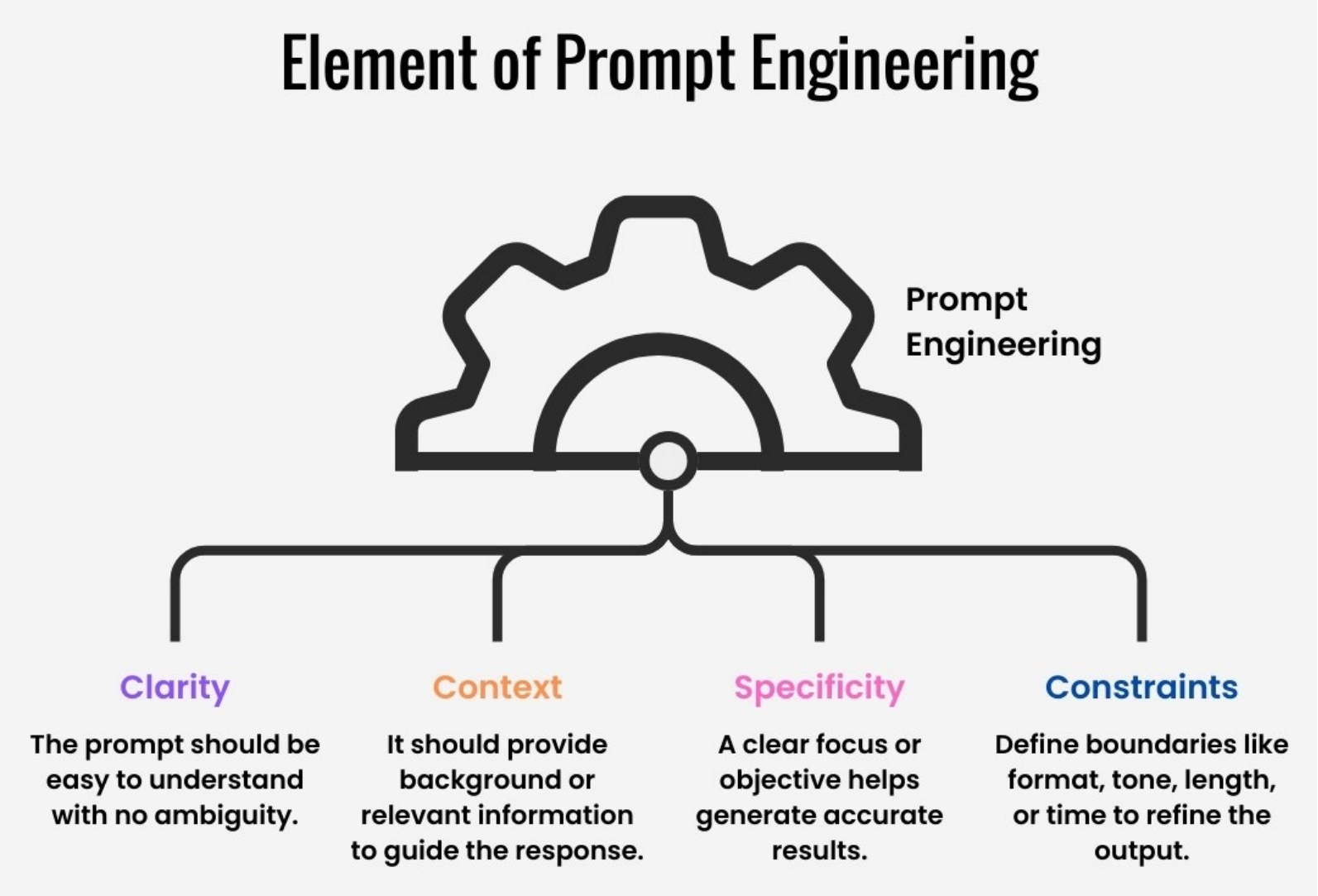
RETRIEVAL: FILTERED VECTOR SEARCH



Filtered Vector Search: filtering to refine the results based on specific criteria

the filtering step is applied before or after the vector search similarity

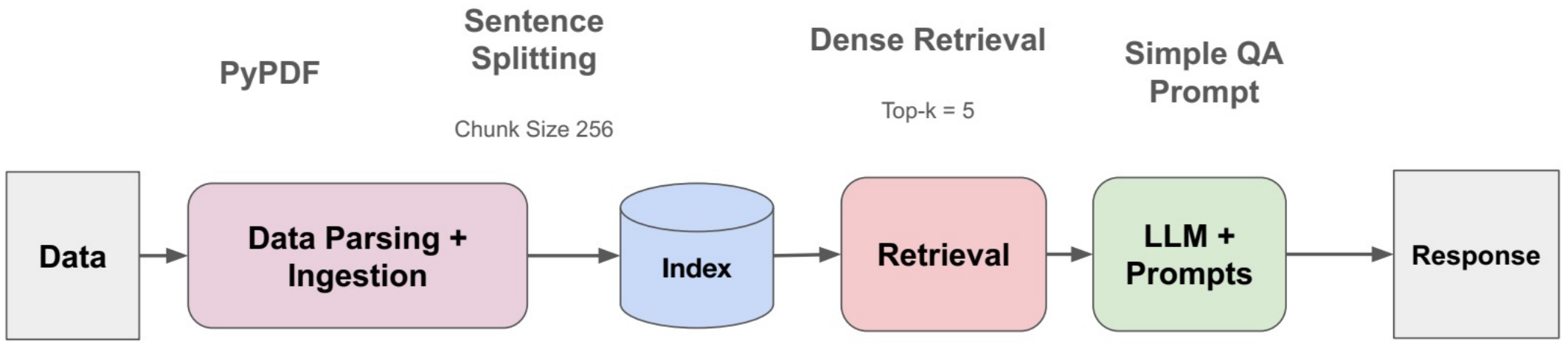
GENERATION


Element of Prompt Engineering
Prompt Engineering
Clarity
The prompt should be easy to understand with no ambiguity.
Context
It should provide background or relevant information to guide the response.
Specificity
A clear focus or objective helps generate accurate results.
Constraints
Define boundaries like format, tone, length, or time to refine the output.

NAÏVE RAG

Naive RAG approaches tend to work well for simple questions over a simple, small set of documents.

- "What are the main risk factors for Tesla?" (over Tesla 2021 10K)
- "What did the author do during his time at YC?" (Paul Graham essay)



CHALLENGES WITH NAIVE RAG (RESPONSE QUALITY)

Bad Retrieval

- ○ Low Precision: Not all chunks in retrieved set are relevant
  - ■ Hallucination + Lost in the Middle Problems
- ○ Low Recall: Now all relevant chunks are retrieved.
  - ■ Lacks enough context for LLM to synthesize an answer
- ○ Outdated information: The data is redundant or out of date.

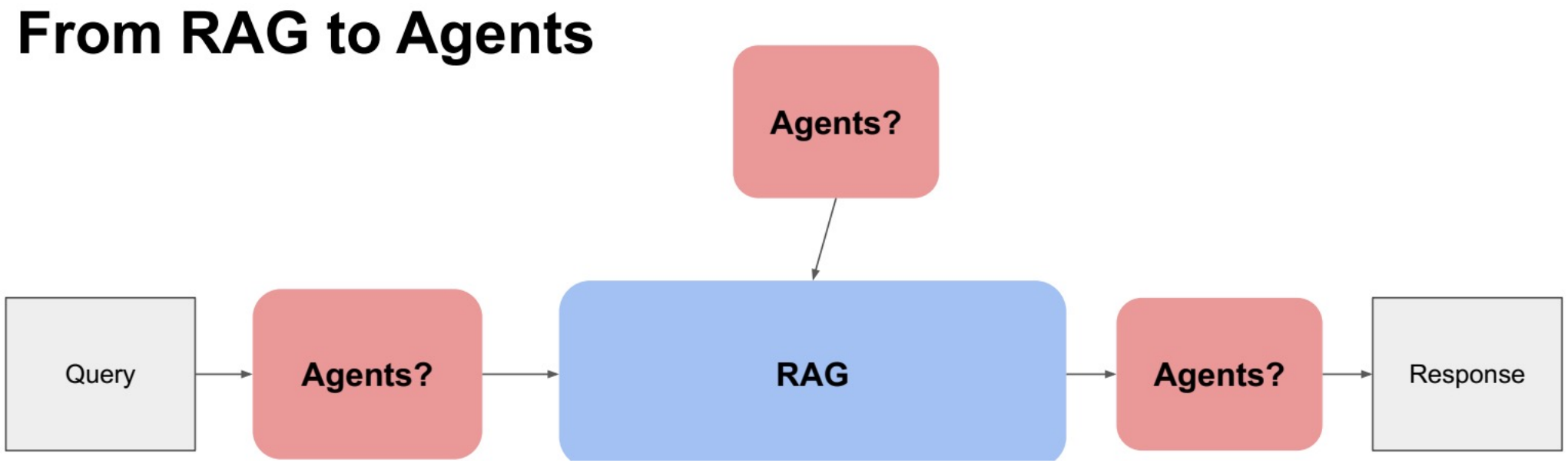
Bad Response Generation

- ○ Hallucination: Model makes up an answer that isn't in the context.
- ○ Irrelevance: Model makes up an answer that doesn't answer the question.
- ○ Toxicity/Bias: Model makes up an answer that's harmful/offensive.

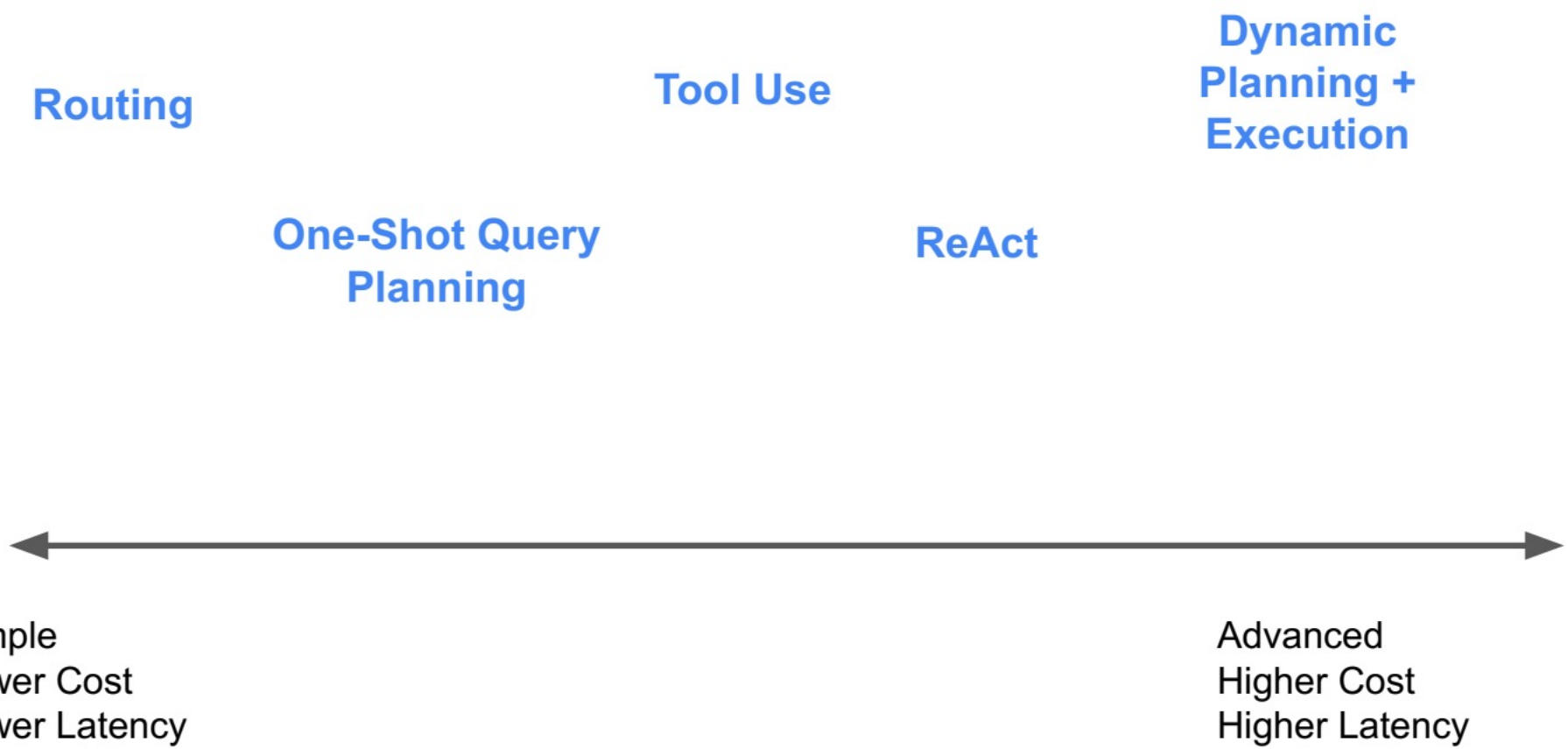
WHAT'S NEXT FOR RAG: AGENTS ?

Agent Definition: Using LLMs for automated reasoning and tool selection

RAG is just one Tool: Agents can decide to use RAG with other tools



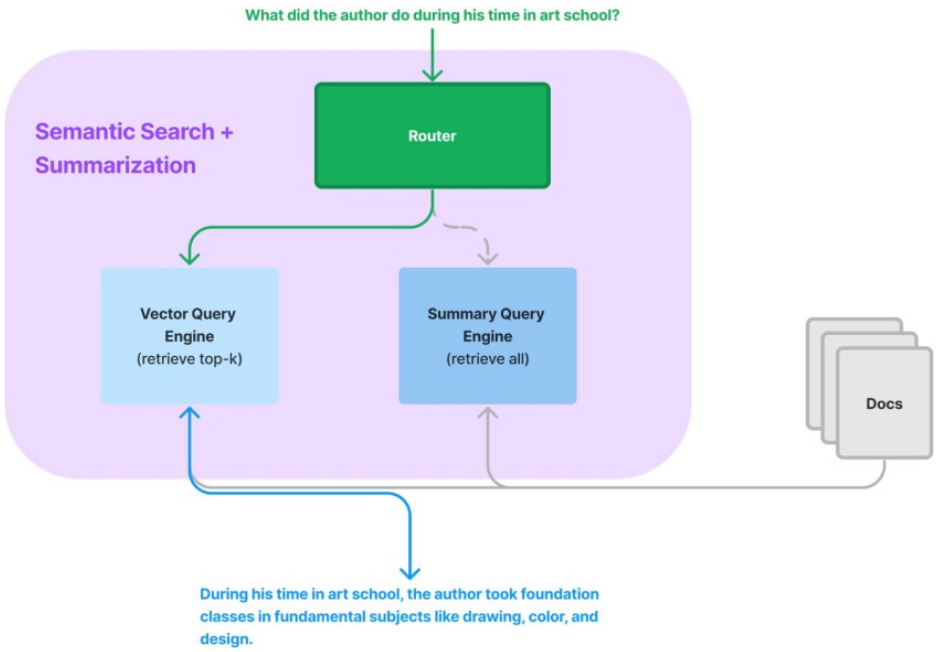
From Simple to Advanced Agents



Routing

Simplest form of agentic reasoning.

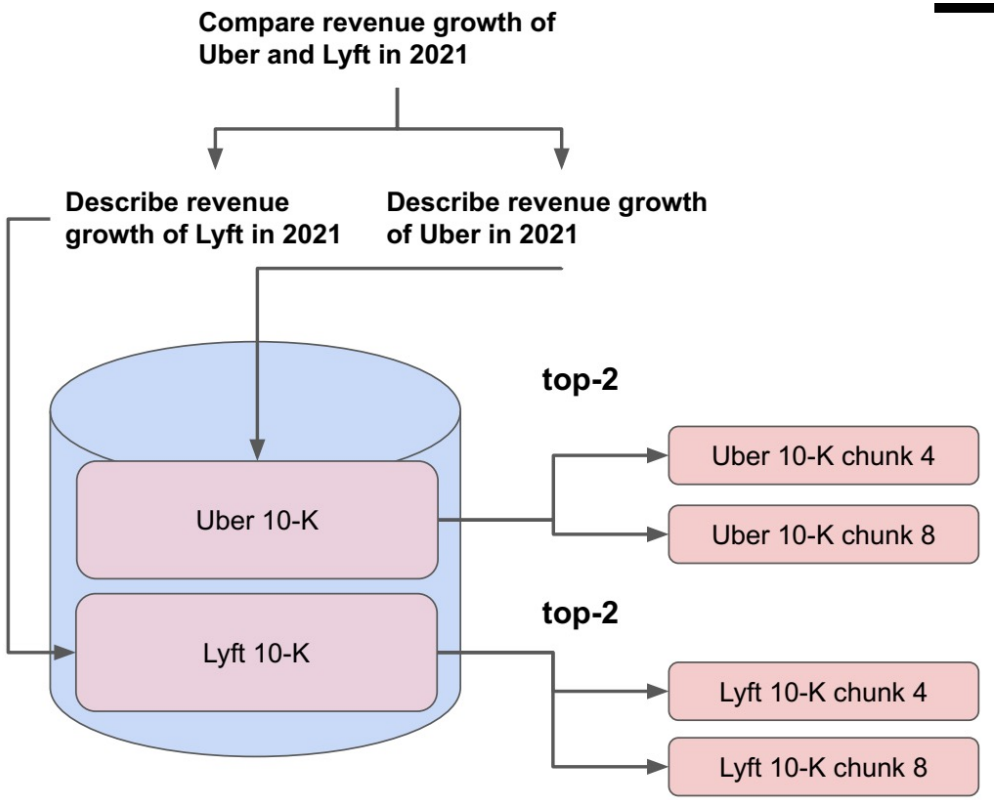
Given user query and set of choices, output subset of choices to route query to



One shot query Planning

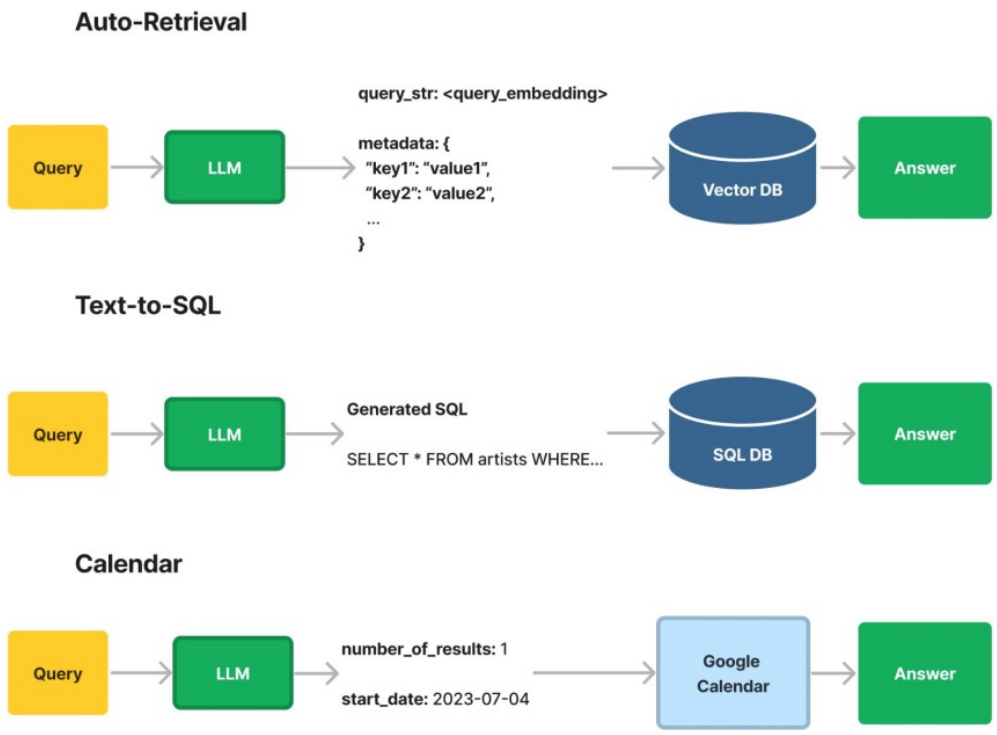
Break down query into parallelizable sub-queries.

Each sub-query can be executed against any set of RAG pipelines

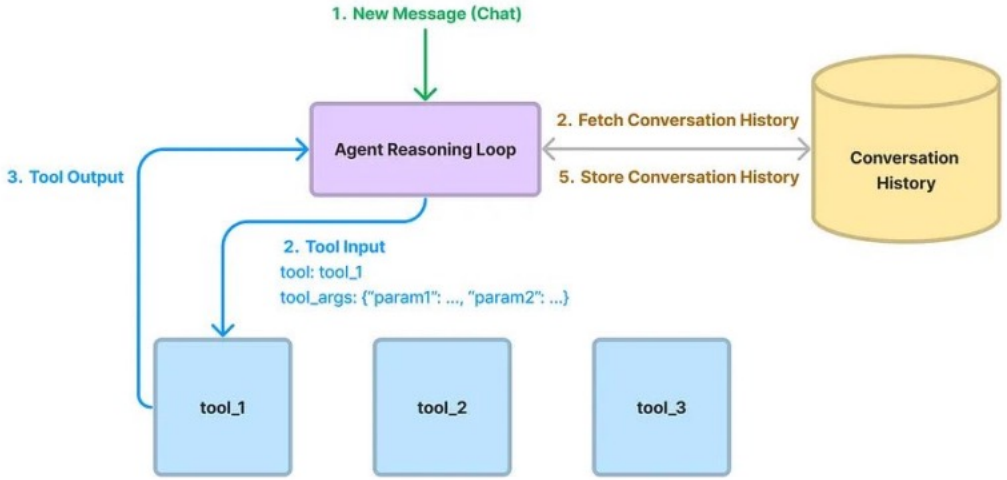


Tool Use

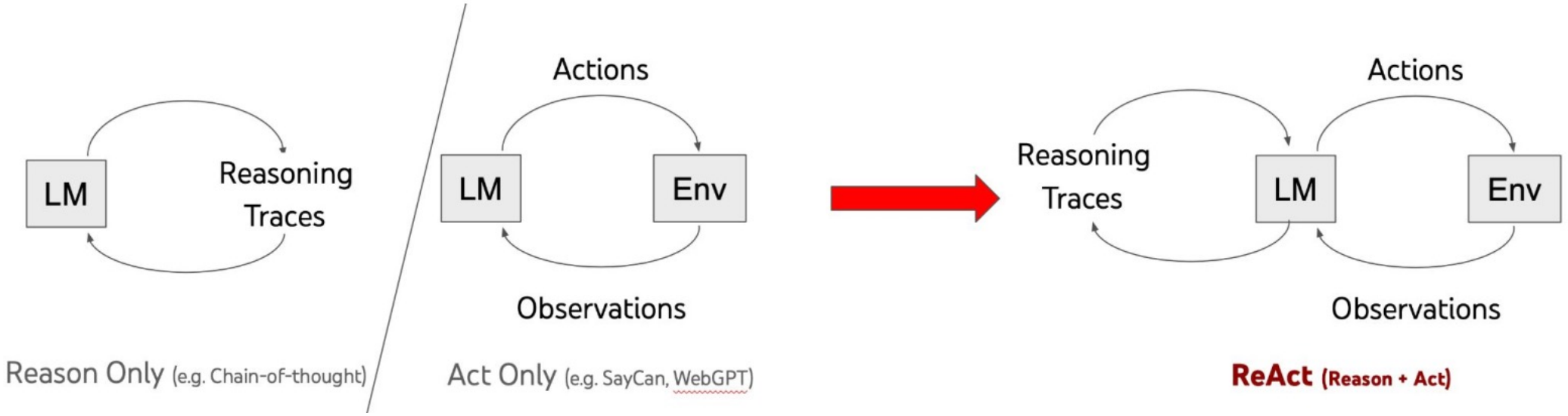
- Use an LLM to call an API
- Infer the parameters of that API



- How can an agent tackle sequential multi-part problems? -> · Let's make it loop
  - ReAct Agent (any LLM)
  - OpenAI Agent (only OAI)
- How can an agent maintain state over time? -> · Let's add basic memory



React: Reasoning + Acting with LLMs



React: Add a loop around Query decomposition + tool use

```
[ ]: response = agent.chat(
         "Tell me about the demographics of Houston, and compare that with the demographics of Chicago"
     )
```

```
Added user message to memory: Tell me about the demographics of Houston, and compare that with the demographics o
f Chicago
=== Calling Function ===
Calling function: vector_tool_Houston with args: {
  "input": "demographics"
}
Got output: Houston has a population of 2,304,580 according to the 2020 U.S. census. In 2017, the estimated popul
ation was 2,312,717, and in 2018 it was 2,325,502. The city has a diverse demographic makeup, with a significant
number of undocumented immigrants residing in the Houston area, comprising nearly 9% of the city's metropolitan p
opulation in 2017. The age distribution in Houston includes a significant number of individuals under 15 and betw
een the ages of 20 to 34. The median age of the city is 33.4. The city has a mix of homeowners and renters, with
an estimated 42.3% of Houstonians owning housing units. The median household income in 2019 was $52,338, and 20.
1% of Houstonians lived at or below the poverty line.
========================

=== Calling Function ===
Calling function: vector_tool_Chicago with args: {
  "input": "demographics"
}
Got output: Chicago experienced rapid population growth during its first hundred years, becoming one of the faste
st-growing cities in the world. From its founding in 1833 with fewer than 200 people, the population grew to over
4,000 within seven years. By 1890, the population had surpassed 1 million, making Chicago the fifth-largest city
in the world at the time. The city's population continued to grow, reaching its highest recorded population of 3.
6 million in 1950. However, in the latter half of the 20th century, Chicago's population declined, dropping to un
der 2.7 million by 2010. The city experienced a rise in population for the 2000 census, followed by a decrease in
2010, and then another increase for the 2020 census. According to U.S. census estimates as of July 2019, the larg
est racial or ethnic groups in Chicago are non-Hispanic White (32.8%), Blacks (30.1%), and Hispanics (29.0%). Add
itionally, Chicago has the third-largest LGBT population in the United States, with an estimated 7.5% of the adul
t population identifying as LGBTQ in 2018.
========================
```

React: Superset of query planning + routing capabilities

```
[ ]: response = agent.chat(
         "Tell me about the demographics of Houston, and compare that with the demographics of Chicago"
     )
```

```
Added user message to memory: Tell me about the demographics of Houston, and compare that with the demographics o
f Chicago
=== Calling Function ===
Calling function: vector_tool_Houston with args: {
  "input": "demographics"
}
Got output: Houston has a population of 2,304,580 according to the 2020 U.S. census. In 2017, the estimated popul
ation was 2,312,717, and in 2018 it was 2,325,502. The city has a diverse demographic makeup, with a significant
number of undocumented immigrants residing in the Houston area, comprising nearly 9% of the city's metropolitan p
opulation in 2017. The age distribution in Houston includes a significant number of individuals under 15 and betw
een the ages of 20 to 34. The median age of the city is 33.4. The city has a mix of homeowners and renters, with
an estimated 42.3% of Houstonians owning housing units. The median household income in 2019 was $52,338, and 20.
1% of Houstonians lived at or below the poverty line.
========================

=== Calling Function ===
Calling function: vector_tool_Chicago with args: {
  "input": "demographics"
}
Got output: Chicago experienced rapid population growth during its first hundred years, becoming one of the faste
st-growing cities in the world. From its founding in 1833 with fewer than 200 people, the population grew to over
4,000 within seven years. By 1890, the population had surpassed 1 million, making Chicago the fifth-largest city
in the world at the time. The city's population continued to grow, reaching its highest recorded population of 3.
6 million in 1950. However, in the latter half of the 20th century, Chicago's population declined, dropping to un
der 2.7 million by 2010. The city experienced a rise in population for the 2000 census, followed by a decrease in
2010, and then another increase for the 2020 census. According to U.S. census estimates as of July 2019, the larg
est racial or ethnic groups in Chicago are non-Hispanic White (32.8%), Blacks (30.1%), and Hispanics (29.0%). Add
itionally, Chicago has the third-largest LGBT population in the United States, with an estimated 7.5% of the adul
t population identifying as LGBTQ in 2018.
========================
```

LLM Compiler

- An agent compiler for parallel multi-function planning + execution
- Plan out steps beforehand, and replan as necessary

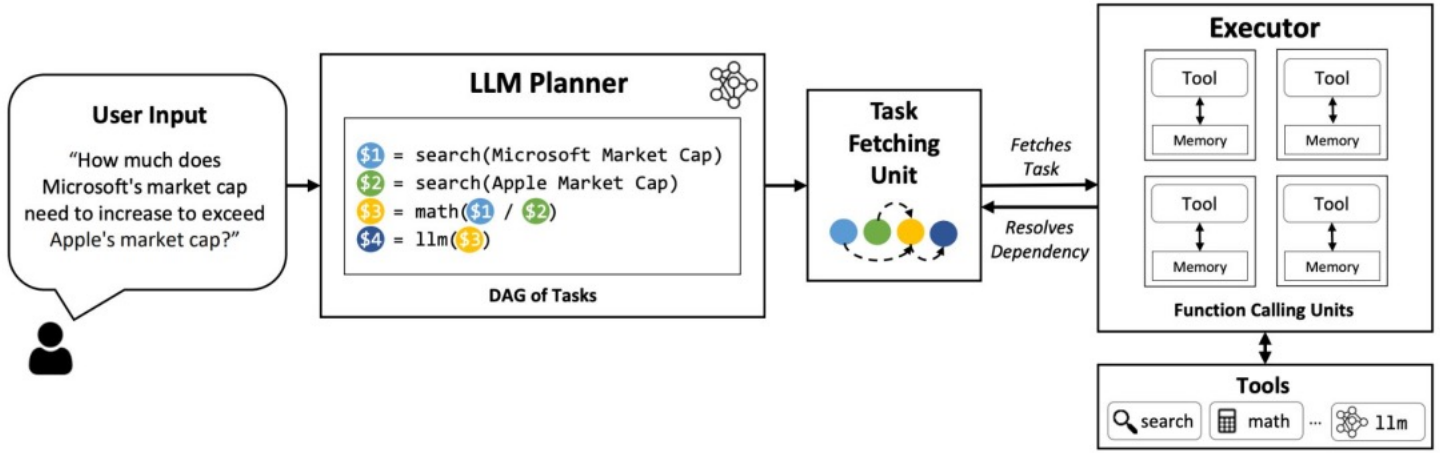


Figure 2: Overview of the `LLMCompiler` framework: the workflow from initial user input to task execution. Beginning with user input, the LLM Planner generates a sequence of tasks with their inter-dependencies. These tasks are then dispatched by the Task Fetching Unit to the Executor based on their dependencies, thus allowing for their parallel executions. For instance, in this example, Task `$1` and `$2` are fetched together for parallel execution of two independent search tasks. After each task is performed, the results (i.e., observations) are forwarded back to the Task Fetching Unit to unblock the dependent tasks after replacing their placeholder variables (e.g., the variable `$1` and `$2` in Task `$3`) with actual values. Once all tasks have been executed, the final answer is delivered to the user.

```
[17]: response = agent.chat(
          "Is the climate of Chicago or Seattle better during the wintertime?"
      )
      print(str(response))
```

```
> Running step f8fdf4cb-9dde-4aba-996d-edbcee53c4c2 for task 20df27d7-cc27-4311-bb57-b1a6f4ad5799.
> Step count: 0
> Plan: 1. vector_tool_Chicago("climate during wintertime")
2. vector_tool_Seattle("climate during wintertime")
3. join()<END_OF_PLAN>
Ran task: vector_tool_Seattle. Observation: During wintertime, Seattle experiences cool, wet conditions. Extreme cold temperatures, below about 15 °F or −9 °C, are rare due to the moderating influence of the adjacent Puget Sound, the greater Pacific Ocean, and Lake Washington. The city is often cloudy due to frequent storms and lows moving in from the Pacific Ocean, and it has many "rain days". However, the rainfall is often a light drizzle.
Ran task: vector_tool_Chicago. Observation: During wintertime, the city experiences relatively cold and snowy conditions. Blizzards can occur, as they did in winter 2011. The normal winter high from December through March is about 36 °F (2 °C). January and February are the coldest months. A polar vortex in January 2019 nearly broke the city's cold record of −27 °F (−33 °C), which was set on January 20, 1985. Measurable snowfall can continue through the first or second week of April. The city's proximity to Lake Michigan tends to keep the lakefront somewhat cooler in summer and less brutally cold in winter than inland parts of the city and suburbs away from the lake. Northeast winds from wintertime cyclones departing south of the region sometimes bring the city lake-effect snow.
Ran task: join. Observation: None
> Thought: Comparing the two climates, Seattle seems to have a milder winter climate than Chicago.
> Answer: Seattle
Seattle
```

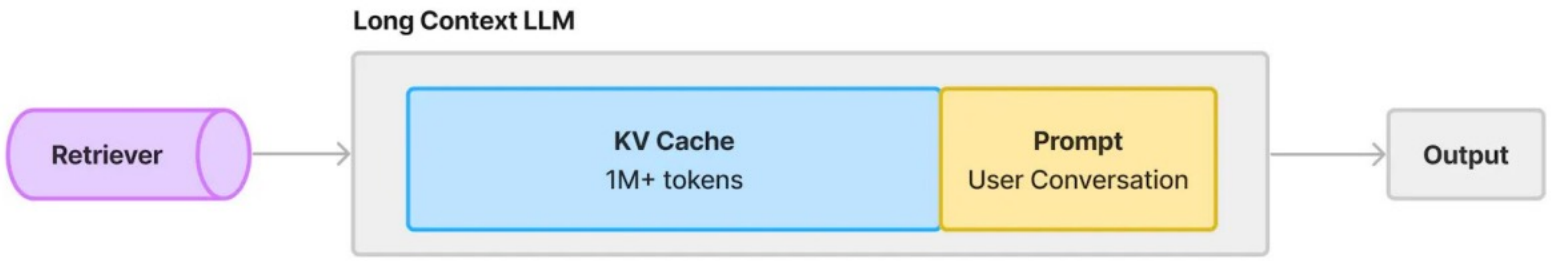
WHAT'S NEXT FOR RAG: LONG CONTEXTS ?

Is RAG Dead?

Gemini 1.5 Pro has a 1-10M context window.

What does this mean for RAG?

https://x.com/Francis_YAO_/status/1759962812229800012?s=20



Yao Fu @Francis_YAO_

Over the last two days after my claim "long context will replace RAG", I have received quite a few criticisms (thanks and really appreciated!) and many of them stand a reasonable point. Here I have gathered the major counterargument, and try to address then one-by-one (feels like a paper rebuttal):

- **RAG is cheap, long context is expensive**. True, but remember, compared to LLM, BERT-small is also cheap, and n-gram is even cheaper, but they are not used today, because we want the model to be smart first, then makes smart models cheaper -- history of AI tells **it is much easier to make smart models cheaper than making cheap model smart** -- when it is cheap, it's never smart.

- **Long context can mix retrieval and reasoning during the whole decoding processing**. RAG only does the retrieval at the very beginning. Typically, given a question, RAG retrieves the paragraphs that is related to the question, then generate. Long-context does the retrieval for every layer and every token. In many cases the model needs to **do on-the-fly per-token interleaved retrieval and reasoning**, and only knows what to retrieve after getting the results of the first reasoning step. Only long-context can do such cases.

- **RAG supports trillion level tokens, long-context is 1M**. True, but there is a natural distribution of the input document, and I tend to believe **most of the cases that requires retrieval is under million level**. For example, imagine a layer working on a case whose input is related legal documents, or a student learning machine learning whose input are three ML books -- does not feel as long as 1B right?

Long Context LLMs will Solve the Following

1. Developers will worry less about tuning chunking algorithms

2. Developers will need to spend less time tuning retrieval and chain-of-thought over single documents

3. Summarization will be easier

4. Personalized memory will be better and easier to build

Some Challenges Remain

1. 10M tokens is not enough for large document corpuses (hundreds ofMB, GB)

2. Embedding models are lagging behind in context length

3. Cost and Latency

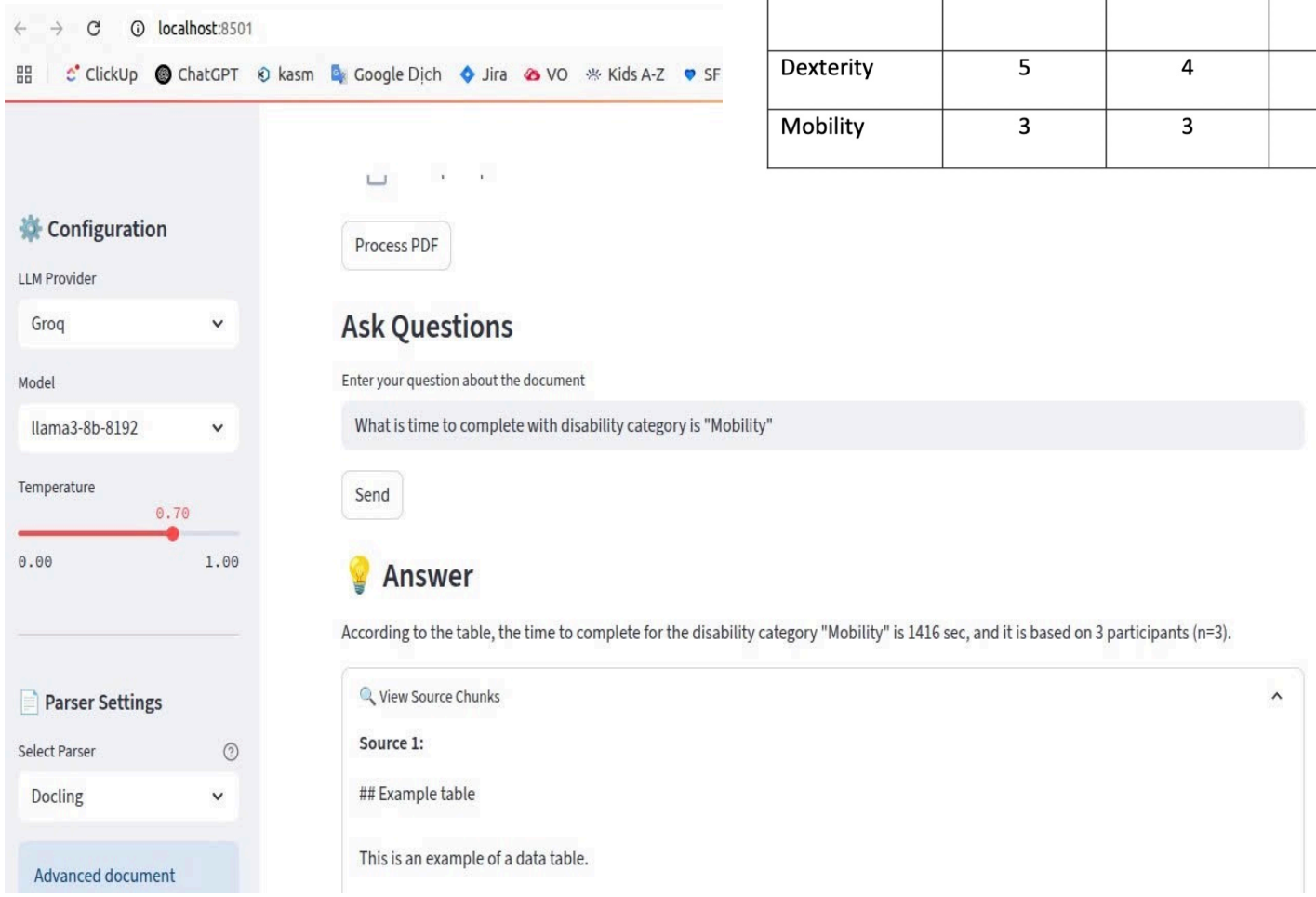
4. A KV Cache takes up a significant amount of GPU memory, and has sequential dependencies

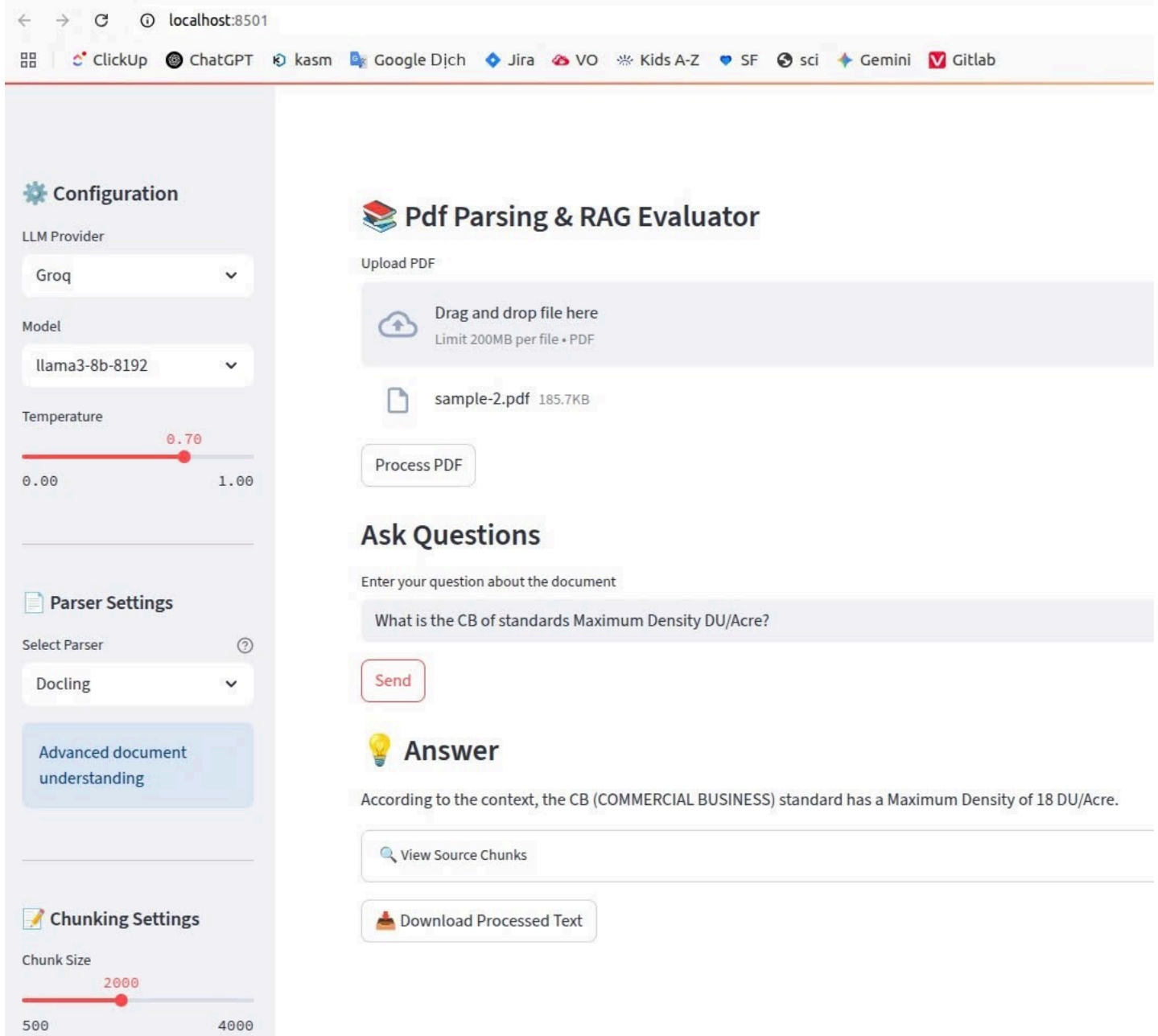
SAMPLE OUTPUTS

Example table

This is an example of a data table.

Disability Category	Participants	Ballots Completed	Ballots Incomplete/ Terminated	Results	
				Accuracy	Time to complete
Blind	5	1	4	34.5%, n=1	1199 sec, n=1
Low Vision	5	2	3	98.3% n=2 (97.7%, n=3)	1716 sec, n=3 (1934 sec, n=2)
Dexterity	5	4	1	98.3%, n=4	1672.1 sec, n=4
Mobility	3	3	0	95.4%, n=3	1416 sec, n=3





21A.25.040 Densities and dimensions – Commercial zones.

A. Commercial Zones.

	ZONES	COMMERCIAL		
		NEIGHBORHOOD BUSINESS	COMMUNITY BUSINESS	OFFICE
STANDARDS		NB	CB	O
Maximum Density DU/Acre		8 du/ac (1)	18 du/ac (1)	18 du/ac (1)
Minimum Lot Area				
Maximum Lot Depth/Width Ratio			10 ft	10 ft
Minimum Street Setback		10 ft (2)	10 ft (2)	10 ft
Minimum Interior Setback (4)		20 ft (5)	20 ft (5)	20 ft (5)
Base Height (7)		35 ft 45 ft	35 ft 60 ft	45 ft 60 ft

Screenshot

Thank You

Feeling gratitude and not expressing it is like wrapping a present and not giving it.