MMDocRAG: Benchmarking Retrieval-Augmented Multimodal Generation for Document QA

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, Yong Liu

Presented by **Noah's Ark Lab, Huawei**

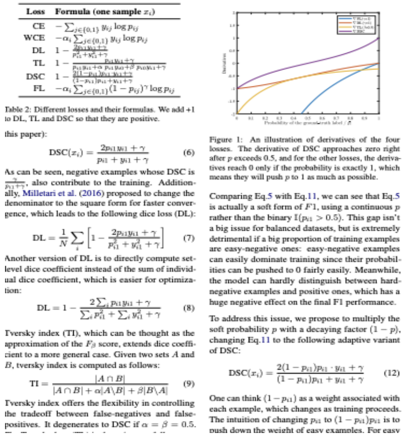
NOAH'S ARK LAB

HUAWEI

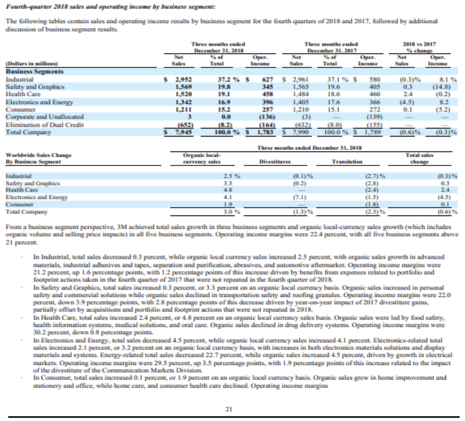
SCAN ME

To download benchmark resources!

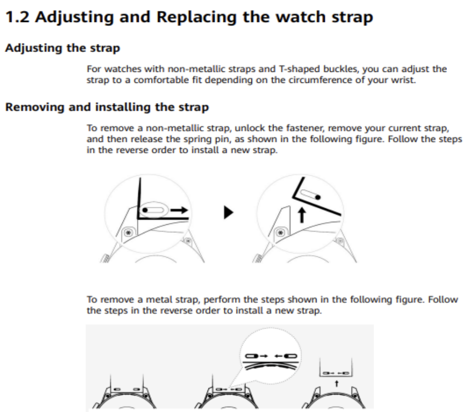
What is Multimodal Document?



Academic Papers

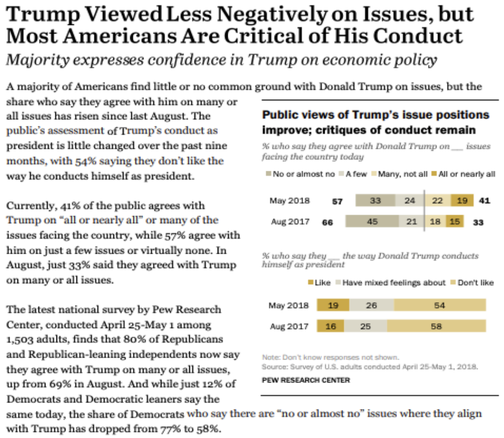


Financial Report

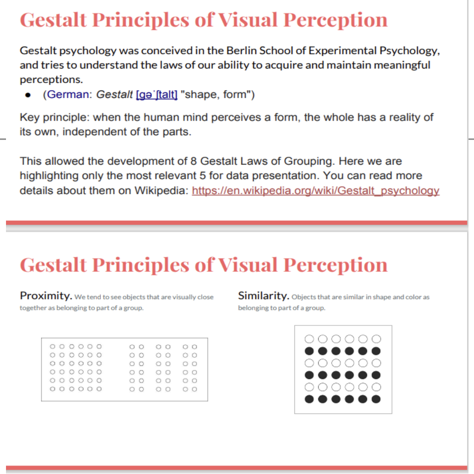


Guidebook

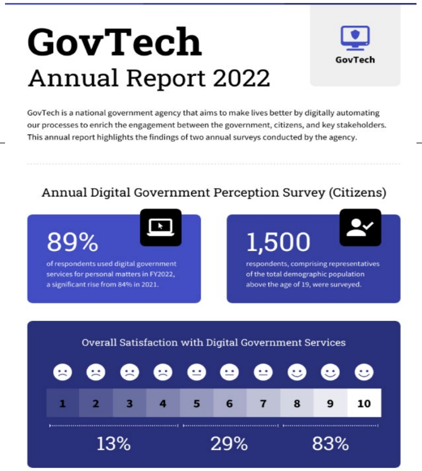
Brochure



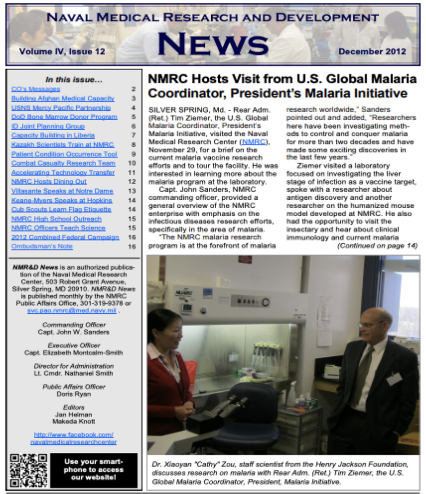
Report



Slides

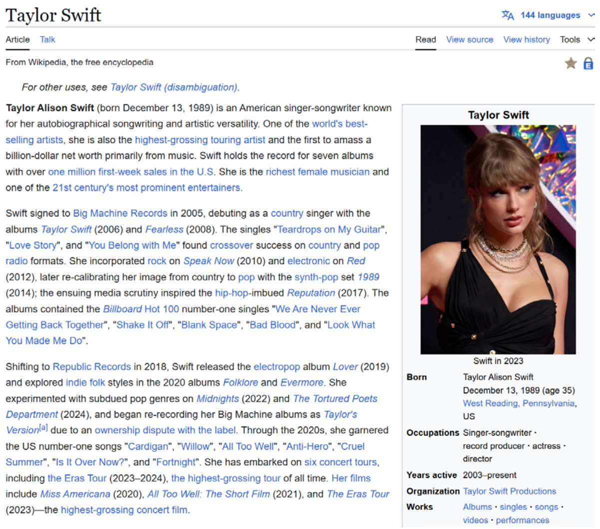


Gov Reports

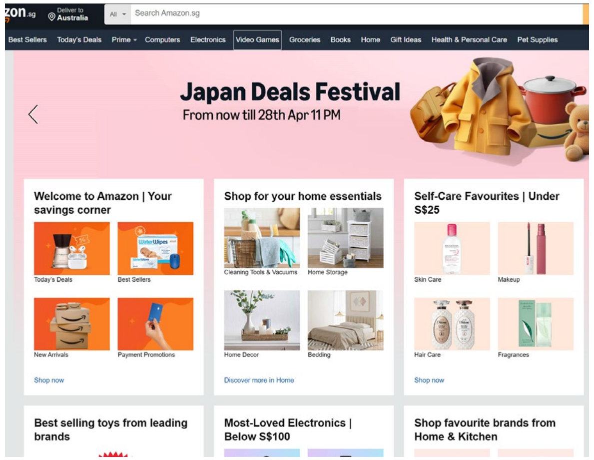


News

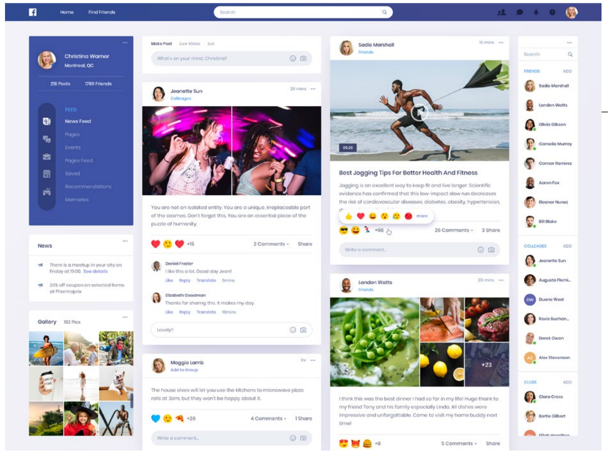
What is Multimodal Document (extended)?



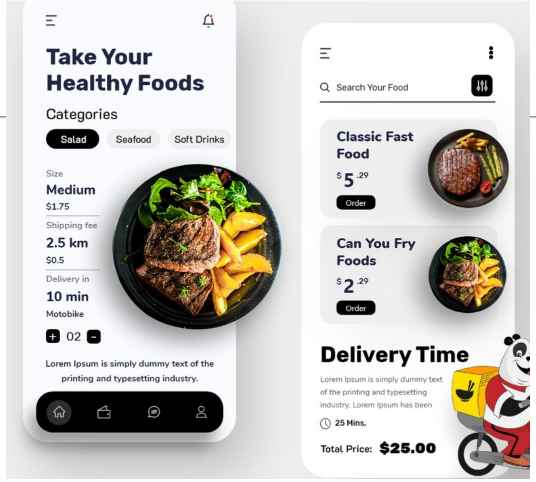
Wiki Pages



E-commerce shopping



Social Media Posts



E-menu

Multimodal Document:

- *Rich texts*
- *Rich visual elements*
- *Complex layouts*
- *Interleaved modalities*
  - *Text*
  - *Equations*
  - *Figures*
  - *Tables*
  - *Charts*
  - *Infographics*
  - *…*

Why Doc Retrieval and Multimodal Generation?

Introduce me about Trump family?

① **AIGC text:** hallucinated, no visual aid.

Donald Trump - The 45&46th President of USA…
Melania Trump - Donald Trump's second wife…
Barron Trump - Donald Trump's youngest children..
Donald Trump Jr. - Trump's eldest child with Ivana…
Ivanka Trump - Trump's second child with Ivana…
Eric Trump - Trump's first child Marla Maples…
Tiffany Trump - Trump's second child with Marla Maples...

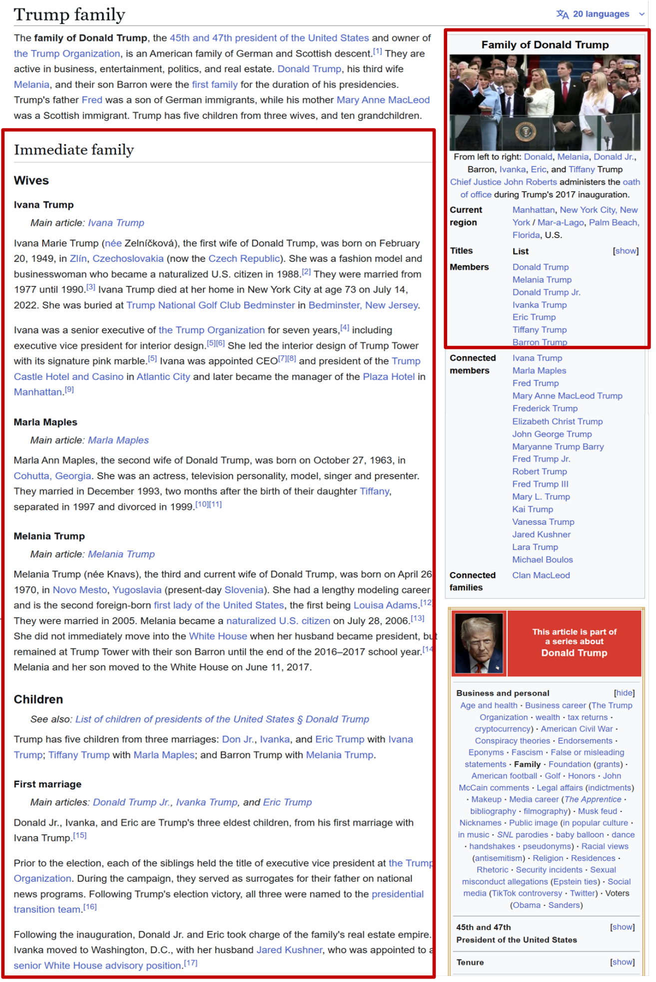
② **AIGC image:** hallucinated, no explanation.

③ **Multimodal RAG:** reliable, traceable, visualizations, and explanations

Donald Trump - The 45&47th President of USA…
Melania Trump - Donald Trump's third wife…
Barron Trump - Donald Trump's youngest children..
Donald Trump Jr. - Trump's eldest child with Ivana…
Ivanka Trump - Trump's second child with Ivana…
Eric Trump - Trump's third child with Ivana…
Tiffany Trump - Trump's only child w/ Marla Maples…

From left to right: Donald, Melania, Donald Jr., Barron, Ivanka, Eric, and Tiffany Trump

Retrieved Docs



Trump family

20 languages

The **family of Donald Trump**, the 45th and 47th president of the United States and owner of the Trump Organization, is an American family of German and Scottish descent.[1] They are active in business, entertainment, politics, and real estate. Donald Trump, his third wife Melania, and their son Barron were the first family for the duration of his presidencies. Trump's father Fred was a son of German immigrants, while his mother Mary Anne MacLeod was a Scottish immigrant. Trump has five children from three wives, and ten grandchildren.

Immediate family

Wives

Ivana Trump

Main article: Ivana Trump

Ivana Marie Trump (née Zelníčková), the first wife of Donald Trump, was born on February 20, 1949, in Zlín, Czechoslovakia (now the Czech Republic). She was a fashion model and businesswoman who became a naturalized U.S. citizen in 1988.[2] They were married from 1977 until 1990.[3] Ivana Trump died at her home in New York City at age 73 on July 14, 2022. She was buried at Trump National Golf Club Bedminster in Bedminster, New Jersey.

Ivana was a senior executive of the Trump Organization for seven years,[4] including executive vice president for interior design.[5][6] She led the interior design of Trump Tower with its signature pink marble.[5] Ivana was appointed CEO[7][8] and president of the Trump Castle Hotel and Casino in Atlantic City and later became the manager of the Plaza Hotel in Manhattan.[9]

Marla Maples

Main article: Marla Maples

Marla Ann Maples, the second wife of Donald Trump, was born on October 27, 1963, in Cohutta, Georgia. She was an actress, television personality, model, singer and presenter. They married in December 1993, two months after the birth of their daughter Tiffany, separated in 1997 and divorced in 1999.[10][11]

Melania Trump

Main article: Melania Trump

Melania Trump (née Knavs), the third and current wife of Donald Trump, was born on April 26 1970, in Novo Mesto, Yugoslavia (present-day Slovenia). She had a lengthy modeling career and is the second foreign-born first lady of the United States, the first being Louisa Adams.[12] They were married in 2005. Melania became a naturalized U.S. citizen on July 28, 2006.[13] She did not immediately move into the White House when her husband became president, but remained at Trump Tower with their son Barron until the end of the 2016–2017 school year.[14] Melania and her son moved to the White House on June 11, 2017.

Children

See also: List of children of presidents of the United States § Donald Trump

Trump has five children from three marriages: Don Jr., Ivanka, and Eric Trump with Ivana Trump; Tiffany Trump with Marla Maples; and Barron Trump with Melania Trump.

First marriage

Main articles: Donald Trump Jr., Ivanka Trump, and Eric Trump

Donald Jr., Ivanka, and Eric are Trump's three eldest children, from his first marriage with Ivana Trump.[15]

Prior to the election, each of the siblings held the title of executive vice president at the Trump Organization. During the campaign, they served as surrogates for their father on national news programs. Following Trump's election victory, all three were named to the presidential transition team.[16]

Following the inauguration, Donald Jr. and Eric took charge of the family's real estate empire. Ivanka moved to Washington, D.C., with her husband Jared Kushner, who was appointed to a senior White House advisory position.[17]

Family of Donald Trump

From left to right: Donald, Melania, Donald Jr., Barron, Ivanka, Eric, and Tiffany Trump Chief Justice John Roberts administers the oath of office during Trump's 2017 inauguration.

Current region	Manhattan, New York City, New York / Mar-a-Lago, Palm Beach, Florida, U.S.
Titles	List [show]
Members	Donald Trump Melania Trump Donald Trump Jr. Ivanka Trump Eric Trump Tiffany Trump Barron Trump
Connected members	Ivana Trump Marla Maples Fred Trump Mary Anne MacLeod Trump Frederick Trump Elizabeth Christ Trump John George Trump Maryanne Trump Barry Fred Trump Jr. Robert Trump Fred Trump III Mary L. Trump Kai Trump Vanessa Trump Jared Kushner Lara Trump Michael Boulos
Connected families	Clan MacLeod

This article is part of a series about Donald Trump

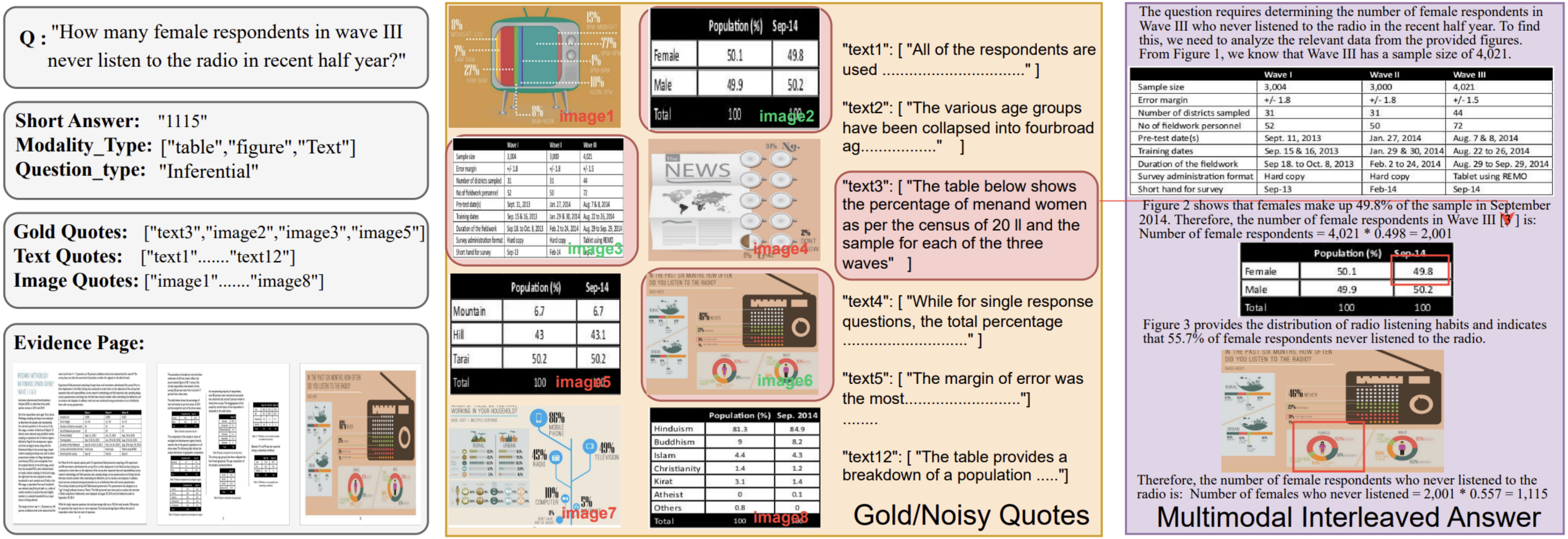
Business and personal [hide]

Age and health · Business career (The Trump Organization · wealth · tax returns · cryptocurrency) · American Civil War · Conspiracy theories · Endorsements · Eponyms · Fascism · False or misleading statements · **Family** · Foundation (grants) · American football · Golf · Honors · John McCain comments · Legal affairs (indictments) · Makeup · Media career (*The Apprentice* · bibliography · filmography) · Musk feud · Nicknames · Public image (in popular culture · in music · *SNL* parodies · baby balloon · dance · handshakes · pseudonyms) · Racial views (antisemitism) · Religion · Residences · Rhetoric · Security incidents · Sexual misconduct allegations (Epstein ties) · Social media (TikTok controversy · Twitter) · Voters (Obama · Sanders)

45th and 47th President of the United States [show]

Tenure [show]

MMDocRAG: Task and Demonstration



QA pair with evidence chains:

- Question/modality type, short answer.
- Evidence page / quote(layout) labels.

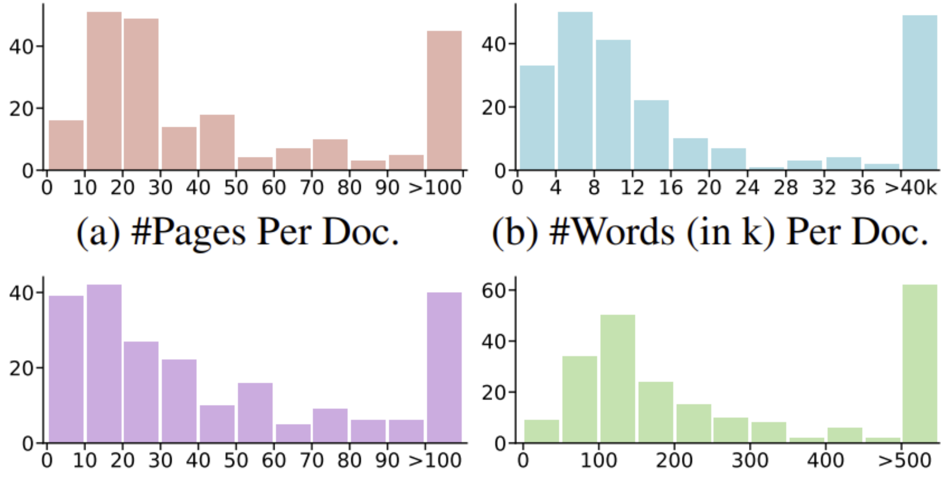
Multimodal Selection and Generation:

- 20 pre-defined quotes (gold mixed w/noisy)
- Interleaved text-image answer

MMDocRAG - More Statistics

Statistic	Number
Documents	222
- Domain Types	10
- Avg./Med./Max. pages per doc	67 / 28 / 844
- Avg./Med./Max. words per doc	33k / 10k / 332k
- Avg./Med./Max. images per doc	63 / 31 / 663
- Avg./Med./Max. texts per doc	536 / 194 / 5k
Total Questions	4,055
- Development / Evaluation split	2,055 / 2,000
- Derived questions	820 (20.2%)
- Newly-annotated questions	3,235 (79.8%)
- Cross-page questions	2,107 (52.0%)
- Multi-image questions	1,590 (39.2%)
- Cross-modal questions	2,503 (61.7%)
(**Question Type**)	
Comparative: 1,456 (35.9%)	Analytical: 488 (12.0%)
Descriptive: 1,256 (31.0%)	Inferential: 75 (1.8%)
Interpretative: 697 (17.2%)	Others: 83 (2.0%)
(**Evidence Modality**)	
Text - 2,457 (60.1%)	Table - 2,677 (66.0%)
Figure - 1,004 (24.8%)	Chart - 636 (15.9%)
All Selected Quotes (Text/Image)	48,618 / 32,071
- Gold Quotes (Text/Image)	4,640 / 6,349
- Noisy Quotes (Text/Image)	43,978 / 25,722
Avg./Med./Max words: question	21.9 / 20 / 73
Avg./Med./Max words: short ans	23.9 / 22 / 102
Avg./Med./Max words: multimodal ans	221.0 / 203 / 768
Avg./Med./Max number of gold quotes	2.7 / 2 / 12

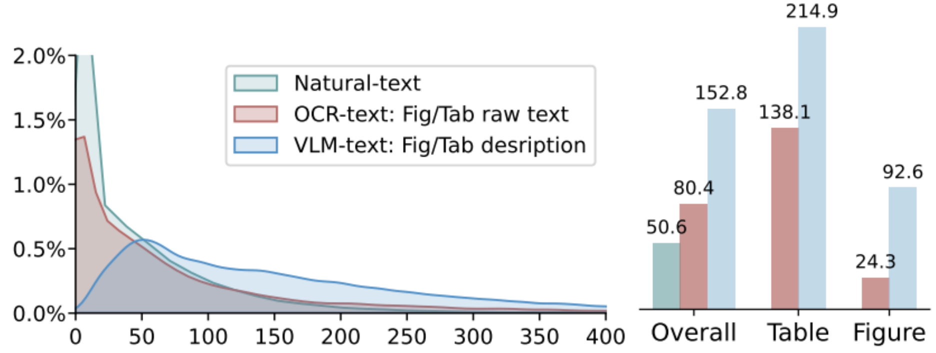
Table 2: Overall Dataset Statistics.



(a) #Pages Per Doc. (b) #Words (in k) Per Doc.

(c) #Image Quotes Per Doc. (d) #Text Quotes Per Doc.

Figure 3: Document Distribution.



(a) Distribution of #words/quote (b) #words/quote

Figure 4: Length Distribution: OCR/VLM-text.

Why choose MMDocRAG?

Benchmarks	Document Domain	Document #Pages	Question #Num	Question Expert	Evi. Loc. Page	Evi. Loc. Quote	Answer Type	Evaluation Metric Evi Loc.	Evaluation Metric Evi Sel.	Evaluation Metric Ans.
MP-DocVQA [70]	Industrial	8,3	46k	✗	✓	✗	TXT	✗	✗	✓
DUDE [33]	Multiple	5.7	24k	✗	✓	✓	TXT	✗	✗	✓
SlideVQA [66]	Slides	20.0	14.5k	✗	✓	✗	TXT	✗	✗	✓
PDF-MVQA [15]	Biomedical	9.6	260k	✗	✓	✓	TXT	✓	✗	✓
MMLongBench-Doc [41]	Multiple	47.5	1,082	✓	✓	✗	TXT	✗	✗	✓
DocBench [85]	Multiple	66.0	1,102	✓	✗	✗	TXT	✗	✗	✓
M3DocVQA [10]	Wikipedia	12.2	2,441	✓	✓	✗	TXT	✓	✗	✓
M-Longdoc [9]	Multiplie	210.8	851	✓	✓	✗	TXT	✓	✗	✓
MMDocIR [16]	Multiple	65.1	1,658	✓	✓	✓	TXT	✓	✗	✗
MuRAR [84]	Webpage	-	300	✓	✗	✗	TXT/TAB/I/V	✗	✗	✓
M^2RAG [42]	Webpage	-	200	✓	✗	✗	TXT/I	✗	✗	✓
MMDocRAG	Multiple	67.0	4,055	✓	✓	✓	TXT/C/TAB/I	✓	✓	✓

High Quality:

- Expert Annotation
- Diverse (10+) domains
- Long docs (67 pages)

All-round Multimodal DocRAG Annotations:

- For retrieval: evidence page/quote labels
- For selection: gold quotes mixed w/ hard negatives
- For generation: multimodal output paradigm

Experiment Setting:

LLM Adaptation:

- Convert multimodal quotes into text format:
  - *OCR-Text*: raw text extracted by OCR tools
  - *VLM-Text*: text descriptions generated by VLM.

Generation Models:

- 33 VLMs: multimodal input quotes
- 27 LLMs: text input quotes
  - VLM-Text or OCR-Text
- 9 Finetuned models
  - 5 LLMs: Qwen-2.5-3,7,14,32,72B.
  - 4 VLMs: Qwen-2.5-VL-3,7B, InternVL-8,9B

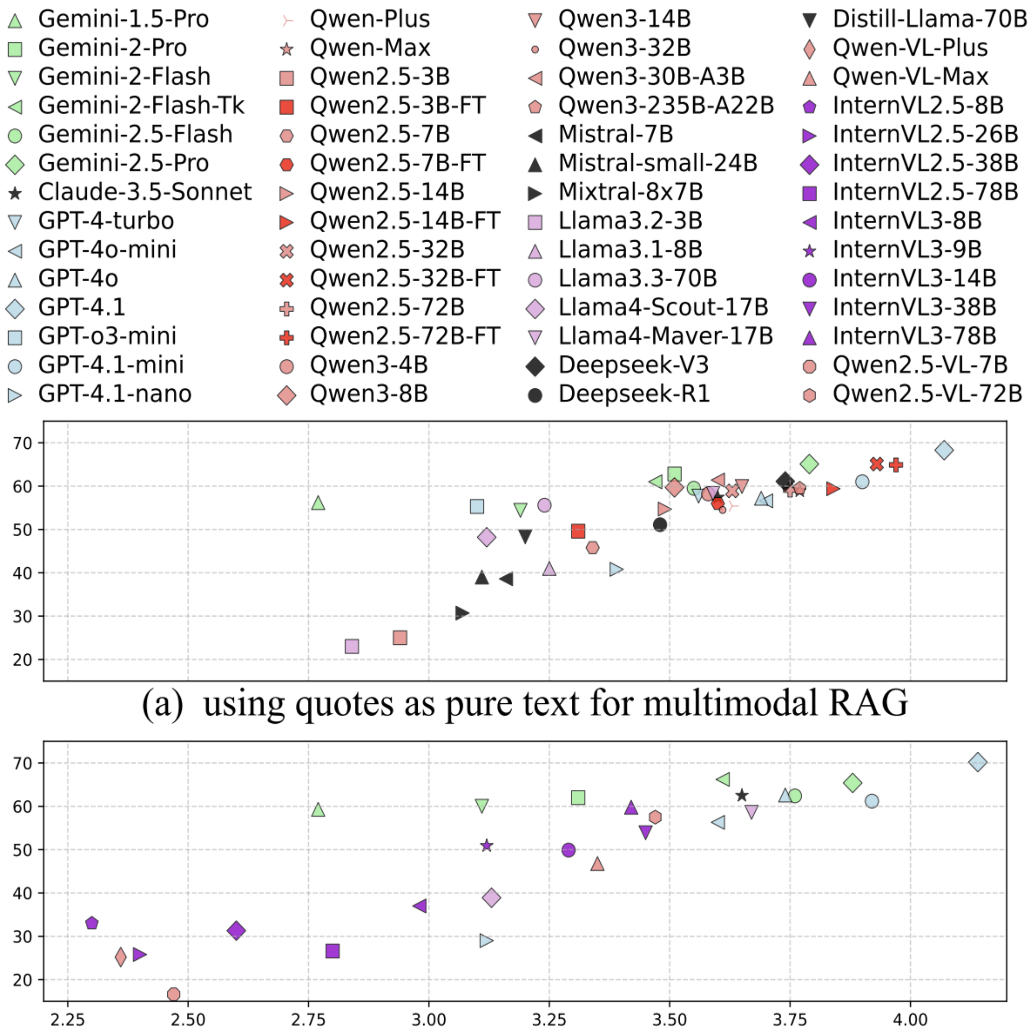
Retrieval Models:

- 14 retrievers:
  - *6 text*: all quotes in text
  - *4 visual*: all quotes in image
  - *4 hybrid*: text/multimodal quotes in text/image, respectively

End-to-end RAG Evaluation:

- Advanced Retrieval system:
  - Ensemble of Single-vector retrievers
  - Query decomposition and reranking
- Generation system:
  - provided with missing gold quotes

Experiment Results: Overall



(a) using quotes as pure text for multimodal RAG

(b) using quotes as interleaved text/image for multimodal RAG

Key Takeaways:

1. GPT-4.1 achieves best results in both quotes selection (70.2) and answer quality (4.14).
2. **Proprietary VLMs** achieve <u>better</u> results using *multimodal quotes vs text quotes*.
3. **Smaller proprietary and open-source models** achieve <u>worse</u> results using *multimodal quotes vs text quotes*.
4. **Smaller VLMs** struggle in multimodal quotes, compared to LLMs using text quotes.
5. **Fine-tuning** can significantly increase the performance.

Fine-tuned Models:

1. LLMs: Qwen-2.5-3,7,14,32,72B.
2. VLMs: Qwen-2.5-VL-3,7B, InternVL-8,9B

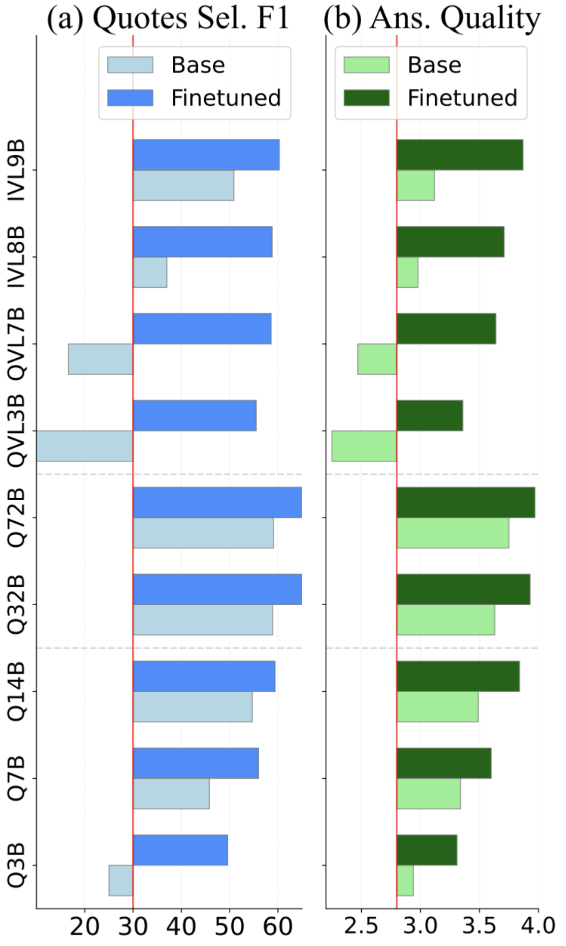


Figure 5: Performance difference: base/finetuned models.

Experiment Results: Multimodal vs Text Quotes

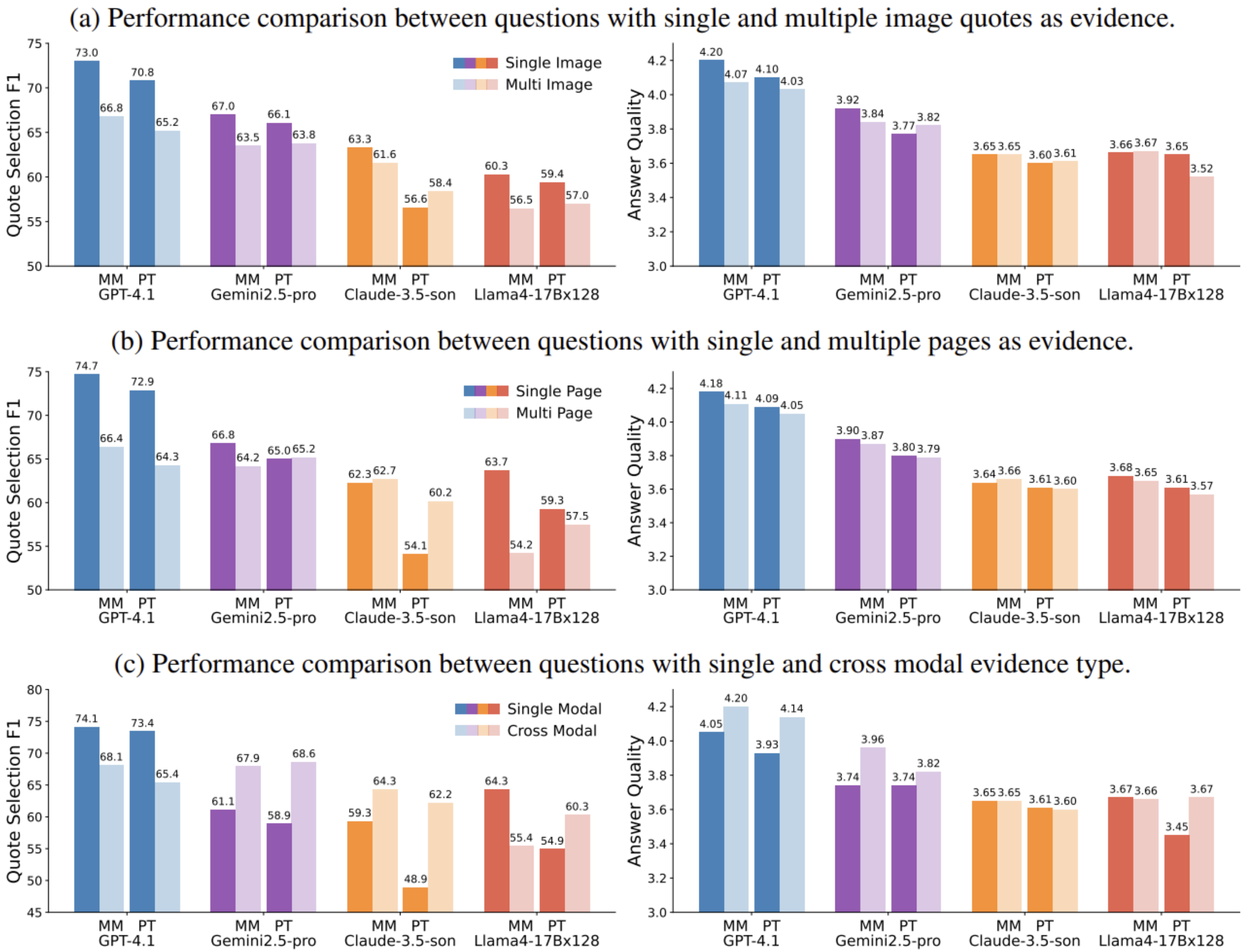
Method Multimodal:MM	Pure-Text:PT	In-token Usage MM	PT	Δ%	Quote Sel. F_1 MM	PT	Δ%	Answer Avg. MM	PT	Δ%
Use the **same VLM** to process both multimodal and pure-text inputs.										
Gemini-1.5-Pro		3.8k	3.6k	-5.3	59.3	56.2	-5.2	2.77	3.03	+9.4
Gemini-2.0-Pro		3.8k	3.6k	-5.3	62.0	62.8	+1.3	3.31	3.51	+6.0
Gemini-2.0-Flash		3.8k	3.6k	-5.3	60.0	54.4	-9.3	3.11	3.19	+2.6
Gemini-2.0-Flash-Think		3.8k	3.6k	-5.3	66.2	61.0	-7.9	3.61	3.47	-3.9
Gemini-2.5-Pro		3.7k	3.6k	-2.7	65.4	65.1	-0.5	3.88	3.79	-2.3
Gemini-2.5-Flash		3.7k	3.6k	-2.7	62.4	59.5	-4.6	3.76	3.55	-5.6
Claude-3.5-Sonnet		7.8k	3.8k	-51.3	62.5	57.4	-8.2	3.65	3.60	-1.4
GPT-4o-mini		8.5k	3.4k	-60.0	56.3	56.6	+0.5	3.60	3.70	+2.8
GPT-4o		6.4k	3.4k	-46.9	62.6	57.2	-8.6	3.74	3.69	-1.3
GPT-4.1-nano		14.2k	3.4k	-76.1	29.0	40.8	+40.7	3.12	3.39	+8.7
GPT-4.1-mini		9.8k	3.4k	-65.3	61.2	61.0	-0.3	3.92	3.90	-0.5
GPT-4.1		6.6k	3.4k	-48.5	70.2	68.3	-2.7	4.14	4.07	-1.7
InternVL3-8B		17.1k	3.6k	-78.9	37.0	48.1	+30.0	3.14	3.19	+1.6
InternVL3-9B		17.2k	4.0k	-76.7	50.9	45.4	-10.8	3.12	3.30	+5.8
InternVL3-14B		17.1k	3.6k	-78.9	49.9	49.9	+0.0	3.29	3.48	+5.8
InternVL3-38B		17.1k	3.6k	-78.9	53.9	55.0	+2.0	3.45	3.61	+4.6
InternVL3-78B		17.1k	3.6k	-78.9	59.8	56.4	-5.7	3.42	3.56	+4.1
Llama4-Scout-17Bx16E		11.6k	3.3k	-71.6	38.9	48.2	+23.9	3.13	3.12	-0.3
Llama4-Mave-17Bx128E		11.6k	3.3k	-71.6	58.6	58.3	-0.5	3.67	3.59	-2.2
Use **separate VLM/LLM** to process multimodal/pure-text inputs, respectively.										
Qw-VL-Plus	Qw-Plus	7.1k	3.6k	-49.3	25.2	55.4	+120	2.36	3.63	+53.8
Qw-VL-Max	Qw-Max	7.1k	3.6k	-49.3	46.8	58.9	+25.9	3.35	3.77	+12.5
QVQ-Max	QwQ-Plus	6.8k	3.6k	-47.1	12.3	59.6	+385	3.36	3.63	+8.0
Qw2.5-VL-7B	Qw2.5-7B	7.1k	3.6k	-49.3	16.6	45.8	+176	2.47	3.34	+35.2
Qw2.5-VL-32B	Qw2.5-32B	7.0k	3.6k	-48.6	36.2	58.9	+62.7	3.73	3.63	-2.7
Qw2.5-VL-72B	Qw2.5-72B	7.1k	3.6k	-49.3	57.5	59.1	+2.8	3.47	3.75	+8.1

Table 4: Using **20 quotes** for multimodal generation. Δ% is calculated by values (PT-MM)/MM in percentage.

Key Takeaways:

1. Multimodal quotes significantly increase token usage
2. Interestingly, Gemini models maintain similar token usage across both modes
3. Gemini, Claude, and GPT models demonstrate superior results in the multimodal setting
4. Qwen models perform better when using pure-text inputs
5. Smaller VLMs, compared to their LLM counterparts, struggle to effectively process long multimodal input sequences.

Experiment Results: Fine-grained Analysis



Key Takeaways:

1. **Single vs. Multiple Image Evidence**: All models consistently achieve higher F1 scores and answer quality when questions require evidence from a single image rather than multiple images.
2. **Single vs. Multiple Page Evidence**: Questions with evidence contained within a single page consistently outperform those requiring multi-page evidence across all models.
3. **Single vs. Cross-Modal Evidence**: Cross-modal evidence preferences vary by model architecture. This reflects differences in how these models handle modality fusion and integration.

Experiment Results: Retrieval and RAG

Model	Retriever	Query	Quote Retrieval Rec. Text	Image	All	Quote Selection F_1 Text	Image	All	Multimodal Answer Quality Bleu	RougeL	LLM-Judge
GPT-4.1	perfect	-	-	-	-	52.0	80.7	70.2	0.157	0.313	4.14
GPT-4.1	BGE	original	34.6	77.8	71.0	34.2	60.8	54.4	0.137	0.299	3.53
GPT-4.1	BGE	clauses	42.1	83.6	78.9	37.9	64.0	57.5	0.141	0.302	3.71
GPT-4.1	multiple	clauses	49.5	86.8	84.9	41.4	65.6	59.9	0.141	0.303	3.79
Gemini2.5-Flash	perfect	-	-	-	-	46.1	76.2	62.4	0.139	0.284	3.76
Gemini2.5-Flash	BGE	original	34.6	77.8	71.0	27.5	55.6	47.7	0.124	0.280	3.21
Gemini2.5-Flash	BGE	clauses	42.1	83.6	78.9	30.9	59.2	50.4	0.125	0.281	3.39
Gemini2.5-Flash	multiple	clauses	49.5	86.8	84.9	34.3	60.3	51.8	0.124	0.281	3.42

Table 7: End-to-end RAG Results.

Retrieval Configuration:

1. **Perfect retriever** (upper bound): Gold quotes mixed w/ noisy quotes
2. **Single retriever** w/ original questions or expanded multi-clause queries.
3. **Multi-retriever** ensemble multiple retrievers with query expansion.

Key Takeaways:

1. **Retrieval-generation correlation**: positive correlation exists between retrieval recall and downstream performance.
2. **Query expansion benefits**: Expanding queries into multi-clause formulations consistently improves retrieval recall.
3. **Multi-retriever robustness**: Ensemble approaches achieve substantially higher recall compared to single retrievers.

THANK YOU

Any Questions?

SCAN ME

To view benchmark resources!