

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



*To download
benchmark
resources!*

What is Multimodal Document?

Loss	Formula (one sample x_i)
CE	$-\sum_{j \in \{0,1\}} y_j \log p_{ij}$
WCE	$-\alpha_i \sum_{j \in \{0,1\}} y_j \log p_{ij}$
DL	$1 - \frac{2p_i y_i + \gamma}{p_i + y_i + \gamma}$
TL	$1 - \frac{2p_i y_i + \gamma}{p_i + y_i + \gamma} \log p_{ij}$
DSC	$-\alpha_i \sum_{j \in \{0,1\}} y_j \log p_{ij}$
FL	$-\alpha_i \sum_{j \in \{0,1\}} y_j \log p_{ij}$

Table 2: Different losses and their formulas. We add +1 to DL, TL, and DSC so that they are positive.

$$DSC(x_i) = \frac{2p_i y_i + \gamma}{p_i + y_i + \gamma} \quad (6)$$

$$DL = \frac{1}{N} \sum_i \left[1 - \frac{2p_i y_i + \gamma}{p_i + y_i + \gamma} \right] \quad (7)$$

Another version of DL is to directly compute set-level dice coefficient instead of the sum of individual dice coefficient, which is easier for optimization:

$$DL = 1 - \frac{2 \sum_i p_i y_i + \gamma}{\sum_i p_i + \sum_i y_i + \gamma} \quad (8)$$

Tversky index (TI), which can be thought as the approximation of the F_β score, extends dice coefficient to a more general case. Given two sets A and B, Tversky index is computed as follows:

$$TI = \frac{|A \cap B|}{|A \cap B| + \alpha |A \setminus B| + \beta |B \setminus A|} \quad (9)$$

Tversky index offers the flexibility in controlling the tradeoff between false-negatives and false-positives. It degenerates to DSC if $\alpha = \beta = 0.5$.

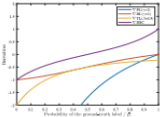


Figure 1: An illustration of derivatives of the four losses. The derivative of DSC approaches zero right after p exceeds 0.5, and for the other losses, the derivatives reach 0 only if the probability is exactly 1, which means they will push p to 1 as much as possible.

Computing Eq 5 with Eq 11, we can see that Eq 5 is actually a soft form of F_1 , using a continuous p rather than the binary $\mathbb{I}(p_i > 0.5)$. This gap isn't a big issue for balanced datasets, but is extremely detrimental if a big proportion of training examples are easy-negative ones: easy-negative examples can easily dominate training since their probabilities can be pushed to 0 fairly easily. Meanwhile, the model can hardly distinguish between hard-negative examples and positive ones, which has a huge negative effect on the final F_1 performance.

To address this issue, we propose to multiply the soft probability p with a decaying factor $(1 - \beta)$, changing Eq 11 to the following adaptive variant of DSC:

$$DSC(x_i) = \frac{2(1 - \beta)p_i y_i + \gamma}{(1 - \beta)p_i + y_i + \gamma} \quad (12)$$

One can think $(1 - \beta)$ as a weight associated with each example, which changes as training proceeds. The intuition of changing p_i to $(1 - \beta)p_i$ is to push down the weight of easy examples. For easy

Fourth-quarter 2018 sales and operating income by business segment

The following tables contain sales and operating income results by business segment for the fourth quarters of 2018 and 2017, followed by additional discussion of business segment results.

	Third quarter ended December 31, 2018			Third quarter ended December 31, 2017			2018 vs 2017 % change		
	Sales	Total	Operating Income	Sales	Total	Operating Income	Sales	Total	Operating Income
Business Segments									
Industrial	\$ 2,952	\$7.2 %	\$ 427	\$ 2,363	\$7.1 %	\$ 500	(0.1)%	6.1%	
Safety and Graphics	1,549	18.8	245	1,385	19.6	242	9.3	(14.8)%	
Health Care	1,539	19.1	408	1,484	18.6	460	2.4	(6.2)%	
Electronics and Energy	1,142	14.9	286	1,485	17.6	368	(4.0)%	8.2	
Consumer	1,211	15.2	257	1,210	15.1	272	0.1	(5.2)%	
Corporate and Unaffiliated	9	0.0	(136)	(43)	(0.0)	(136)			
Elimination of Debt Costs	(823)	(10.2)	(146)	(823)	(10.2)	(146)			
Total Company	\$ 7,924	\$6.8 %	\$ 1,323	\$ 7,327	\$6.7 %	\$ 1,739	(7.2)%	(13.7)%	

	Third quarter ended December 31, 2018			Third quarter ended December 31, 2017			2018 vs 2017 % change		
	Operating Income	Current sales	Disclosures	Operating Income	Current sales	Disclosures	Operating Income	Current sales	Disclosures
Industrial	2.5 %	2.5 %	(0.1)%	(0.1)%	2.5 %	(0.1)%	0.0%	0.0%	0.0%
Safety and Graphics	5.3	5.3	(0.2)	5.3	5.3	(0.2)	0.0%	0.0%	0.0%
Health Care	4.4	4.4	(2.4)	4.4	4.4	(2.4)	0.0%	0.0%	0.0%
Electronics and Energy	4.1	4.1	(1.5)	4.1	4.1	(1.5)	0.0%	0.0%	0.0%
Consumer	1.9	1.9	(0.8)	1.9	1.9	(0.8)	0.0%	0.0%	0.0%
Total Company	3.2 %	3.2 %	(1.3)%	3.2 %	3.2 %	(1.3)%	0.0%	0.0%	0.0%

From a business segment perspective, 3M achieved total sales growth in three business segments and organic local-currency sales growth (which includes organic volume and selling price impacts) in all five business segments. Operating income margins were 22.4 percent, with all five business segments above 21 percent.

- In Industrial, total sales decreased 0.3 percent, while organic local-currency sales increased 2.5 percent, with organic sales growth in advanced materials, industrial adhesives and tapes, separation and purification, abrasives, and automotive aftermarket. Operating income margins were 21.2 percent, up 1.6 percentage points, with 1.2 percentage points of this increase driven by benefits from expenses related to portfolio and foreign exchange gains in the fourth quarter of 2017 that were not repeated in the fourth quarter of 2018.
- In Safety and Graphics, total sales increased 0.1 percent, or 3.3 percent on an organic local-currency basis. Organic sales increased in personal safety and commercial solutions while organic sales declined in transportation safety and roofing granules. Operating income margins were 5.3 percent, down 1.9 percentage points, with 2.8 percentage points of this decrease driven by year-on-year impact of 2017 divestiture gains, partially offset by acquisition and portfolio, and cost cuts. Organic sales declined in drug delivery systems. Operating income margins were 5.3 percent, down 0.8 percentage points.
- In Electronics and Energy, total sales decreased 4.5 percent, while organic local-currency sales increased 4.1 percent. Electronics-related total sales increased 2.1 percent, or 3.3 percent on an organic local-currency basis, with increases in health information systems and display materials and systems. Energy-related total sales decreased 2.7 percent, while organic sales increased 4.5 percent, driven by growth in electrical materials. Operating income margins were 2.5 percent, up 1.5 percentage points, with 1.9 percentage points of this increase related to the impact of the divestiture of the Consumer Markets Division.
- In Consumer, total sales increased 0.1 percent, or 1.5 percent on an organic local-currency basis. Organic sales grew in home improvement and stationary and office, while home care, and consumer health care declined. Operating income margins

Academic Papers

Trump Viewed Less Negatively on Issues, but Most Americans Are Critical of His Conduct

Majority expresses confidence in Trump on economic policy

A majority of Americans find little or no common ground with Donald Trump on issues, but the share who say they agree with him on many or all issues has risen since last August. The public's assessment of Trump's conduct as president is little changed over the past nine months, with 54% saying they don't like the way he conducts himself as president.

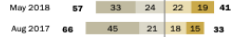
Currently, 41% of the public agrees with Trump on "all or nearly all" or many of the issues facing the country, while 57% agree with him on just a few issues or virtually none. In August, just 33% said they agreed with Trump on many or all issues.

The latest national survey by Pew Research Center, conducted April 25-May 1 among 1,503 adults, finds that 80% of Republicans and Republican-leaning independents now say they agree with Trump on many or all issues, up from 69% in August. And while just 12% of Democrats and Democratic leaners say the same today, the share of Democrats who say there are "no or almost no" issues where they align with Trump has dropped from 77% to 58%.

Public views of Trump's issue positions improve; critiques of conduct remain

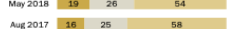
% who say they agree with Donald Trump on ____ issues facing the country today

■ No or almost no ■ A few ■ Many, not all ■ All or nearly all



% who say they ____ the way Donald Trump conducts himself as president

■ Like ■ Have mixed feelings about ■ Don't like



Note: Don't know responses not shown.
Source: Survey of U.S. adults conducted April 25-May 1, 2018.
PEW RESEARCH CENTER

Financial Report

Gestalt Principles of Visual Perception

Gestalt psychology was conceived in the Berlin School of Experimental Psychology, and tries to understand the laws of our ability to acquire and maintain meaningful perceptions.

- (German: *Gestalt* [[ɡəˈtalt](#)] "shape, form")

Key principle: when the human mind perceives a form, the whole has a reality of its own, independent of the parts.

This allowed the development of 8 Gestalt Laws of Grouping. Here we are highlighting only the most relevant for data presentation. You can read more details about them on Wikipedia: https://en.wikipedia.org/wiki/Gestalt_psychology

Gestalt Principles of Visual Perception

Proximity. We tend to see objects that are visually close together as belonging to part of a group.



Similarity. Objects that are similar in shape and color as belonging to part of a group.



Report

Slides

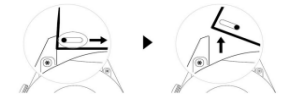
1.2 Adjusting and Replacing the watch strap

Adjusting the strap

For watches with non-metallic straps and T-shape buckles, you can adjust the strap to a comfortable fit depending on the circumference of your wrist.

Removing and installing the strap

To remove a non-metallic strap, unlock the fastener, remove your current strap, and then release the spring pin, as shown in the following figure. Follow the steps in the reverse order to install a new strap.



To remove a metal strap, perform the steps shown in the following figure. Follow the steps in the reverse order to install a new strap.

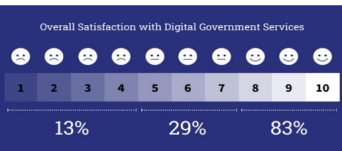


Guidebook

GovTech Annual Report 2022

GovTech is a national government agency that aims to make lives better by digitally automating our processes to enrich the engagement between the government, citizens, and key stakeholders. This annual report highlights the findings of two annual surveys conducted by the agency.

Annual Digital Government Perception Survey (Citizens)



Gov Reports

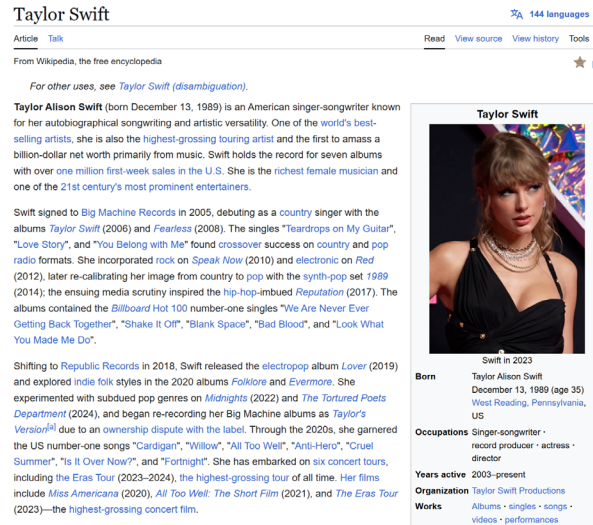


Brochure

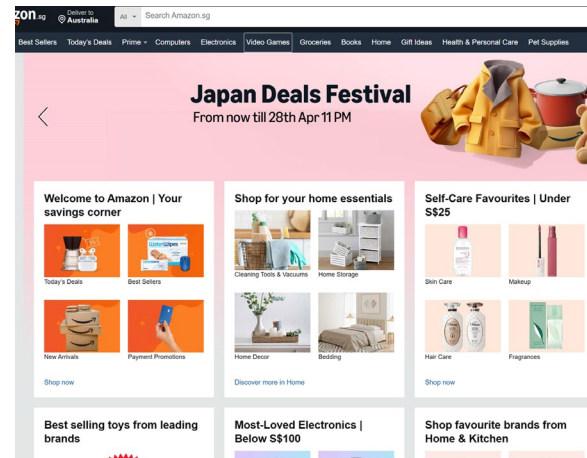


News

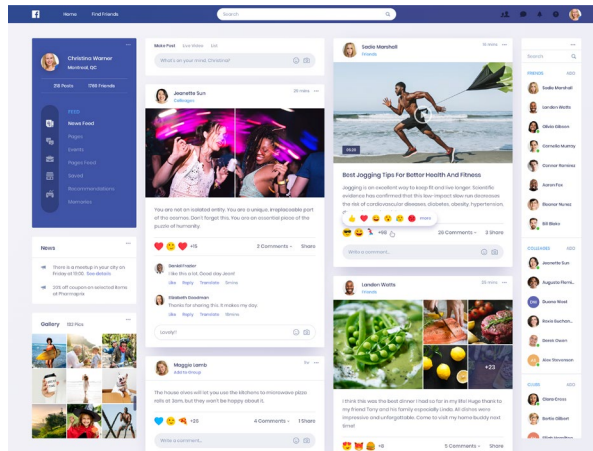
What is Multimodal Document (extended)?



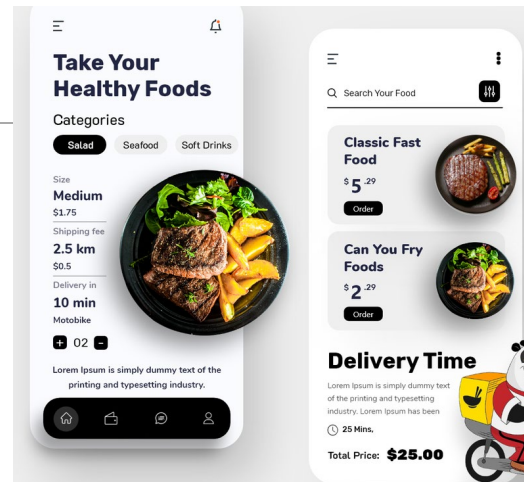
Wiki Pages



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

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.^[6] 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 Knaws), 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

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
Clan MacLeod

Connected families

Clan MacLeod

This article is part of a series about Donald Trump

Business and personal

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) • Music • 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

Tenure

MMDocRAG: Task and Demonstration

Q : "How many female respondents in wave III never listen to the radio in recent half year?"

Short Answer: "1115"

Modality_Type: ["table", "figure", "Text"]

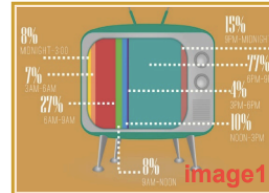
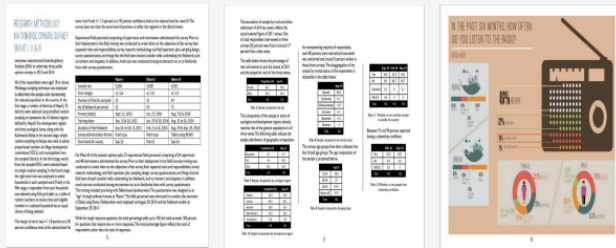
Question_type: "Inferential"

Gold Quotes: ["text3", "image2", "image3", "image5"]

Text Quotes: ["text1"....."text12"]

Image Quotes: ["image1"....."image8"]

Evidence Page:

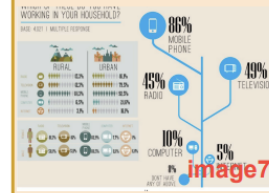
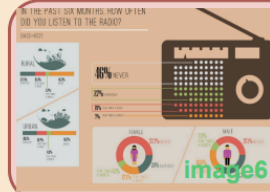


Population (%) Sep-14		
Female	50.1	49.8
Male	49.9	50.2
Total	100	100

	Wave I	Wave II	Wave III
Sample size	3,004	3,000	4,021
Error margin	+/- 1.8	+/- 1.8	+/- 1.5
Number of districts sampled	31	31	44
No of fieldwork personnel	52	50	72
Pre-test date(s)	Sep. 11, 2013	Jan. 27, 2014	Aug. 7 & 8, 2014
Training dates	Sep. 15 & 16, 2013	Jan. 29 & 30, 2014	Aug. 22 to 26, 2014
Duration of the fieldwork	Sep. 18 to Oct. 8, 2013	Feb. 2 to 24, 2014	Aug. 29 to Sep. 29, 2014
Survey administration format	Hard copy	Hard copy	Tablet using REMO
Short hand for survey	Sep-13	Feb-14	Sep-14



Population (%) Sep-14		
Mountain	6.7	6.7
Hill	43	43.1
Tarai	50.2	50.2
Total	100	100



Population (%) Sep. 2014		
Hinduism	81.3	84.9
Buddhism	9	8.2
Islam	4.4	4.3
Christianity	1.4	1.2
Kirat	3.1	1.4
Atheist	0	0.1
Others	0.8	0
Total	100	100

"text1": ["All of the respondents are used"]

"text2": ["The various age groups have been collapsed into four broad ag....."]

"text3": ["The table below shows the percentage of men and women as per the census of 2011 and the sample for each of the three waves"]

"text4": ["While for single response questions, the total percentage"]

"text5": ["The margin of error was the most....."]

"text12": ["The table provides a breakdown of a population"]

Gold/Noisy Quotes

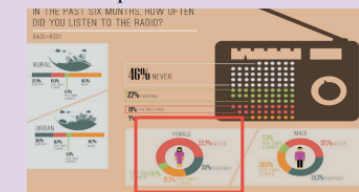
The question requires determining the number of female respondents in Wave III who never listened to the radio in the recent half year. To find this, we need to analyze the relevant data from the provided figures. From Figure 1, we know that Wave III has a sample size of 4,021.

	Wave I	Wave II	Wave III
Sample size	3,004	3,000	4,021
Error margin	+/- 1.8	+/- 1.8	+/- 1.5
Number of districts sampled	31	31	44
No of fieldwork personnel	52	50	72
Pre-test date(s)	Sep. 11, 2013	Jan. 27, 2014	Aug. 7 & 8, 2014
Training dates	Sep. 15 & 16, 2013	Jan. 29 & 30, 2014	Aug. 22 to 26, 2014
Duration of the fieldwork	Sep. 18 to Oct. 8, 2013	Feb. 2 to 24, 2014	Aug. 29 to Sep. 29, 2014
Survey administration format	Hard copy	Hard copy	Tablet using REMO
Short hand for survey	Sep-13	Feb-14	Sep-14

Figure 2 shows that females make up 49.8% of the sample in September 2014. Therefore, the number of female respondents in Wave III is: Number of female respondents = 4,021 * 0.498 = 2,001

Population (%) Sep-14		
Female	50.1	49.8
Male	49.9	50.2
Total	100	100

Figure 3 provides the distribution of radio listening habits and indicates that 55.7% of female respondents never listened to the radio.



Therefore, the number of female respondents who never listened to the radio is: Number of females who never listened = 2,001 * 0.557 = 1,115

Multimodal Interleaved Answer

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.

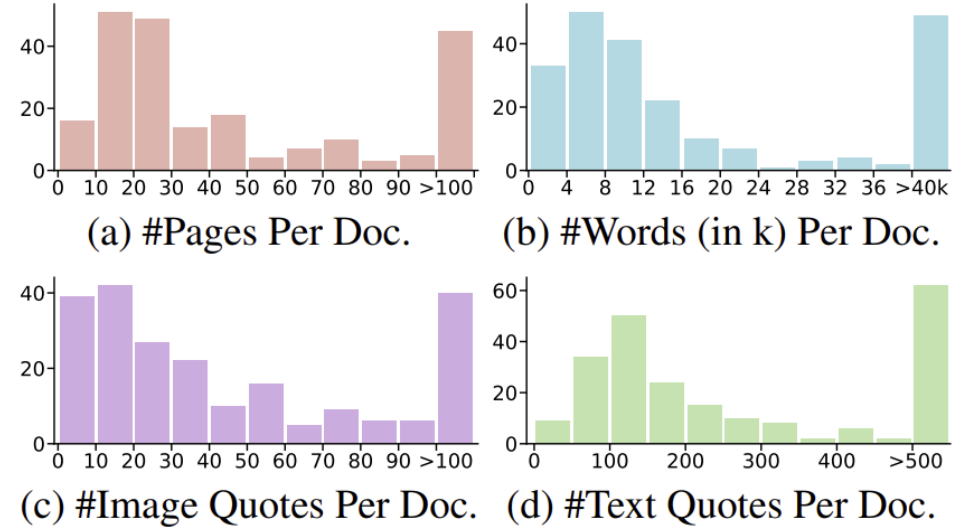
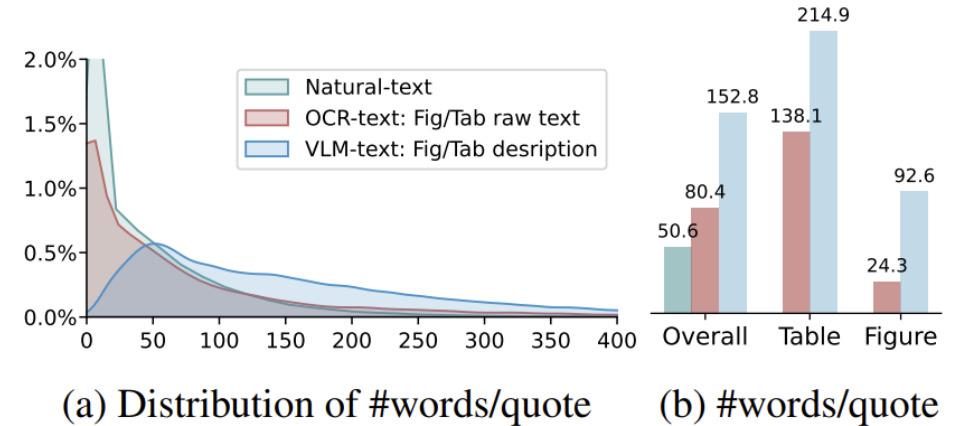


Figure 3: Document Distribution.



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

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

Why choose MMDocRAG?

Benchmarks	Document		Question		Evi. Loc.		Answer Type	Evaluation Metric		
	Domain	#Pages	#Num	Expert	Page	Quote		Evi	Loc.	Evi Sel. 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]	Multiple	210.8	851	✓	✓	✗	TXT	✓	✗	✓
MMDocIR [16]	Multiple	65.1	1,658	✓	✓	✓	TXT	✓	✗	✗
MuRAR [84]	Webpage	-	300	✓	✗	✗	TXT/TAB/I/V	✗	✗	✓
M ² RAG [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.

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

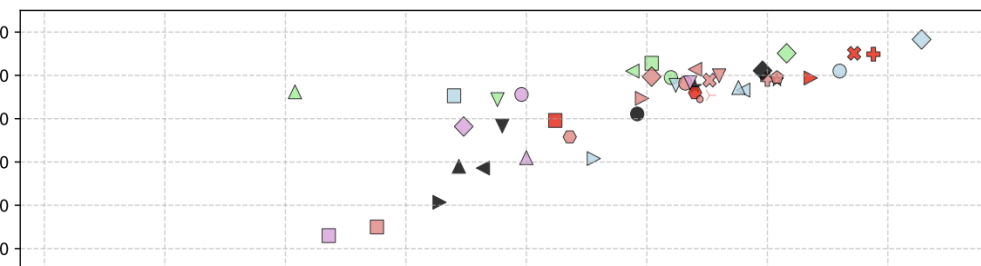
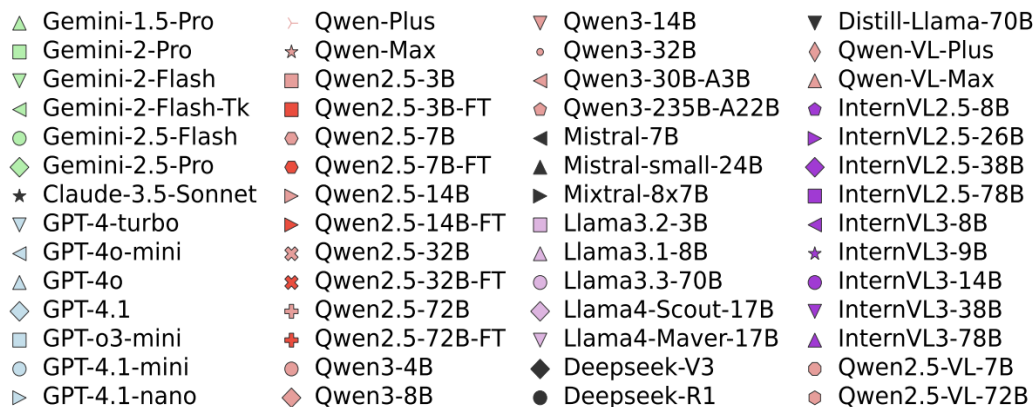
Generation Models:

- 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

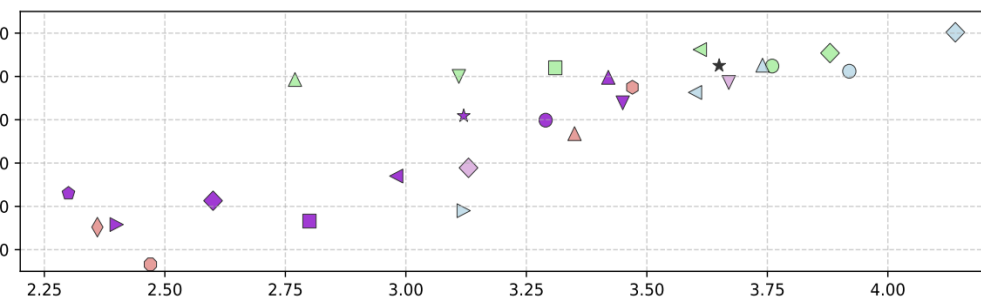
End-to-end RAG Evaluation:

- 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 better results using *multimodal quotes vs text quotes*.
3. **Smaller proprietary and open-source models** achieve worse 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

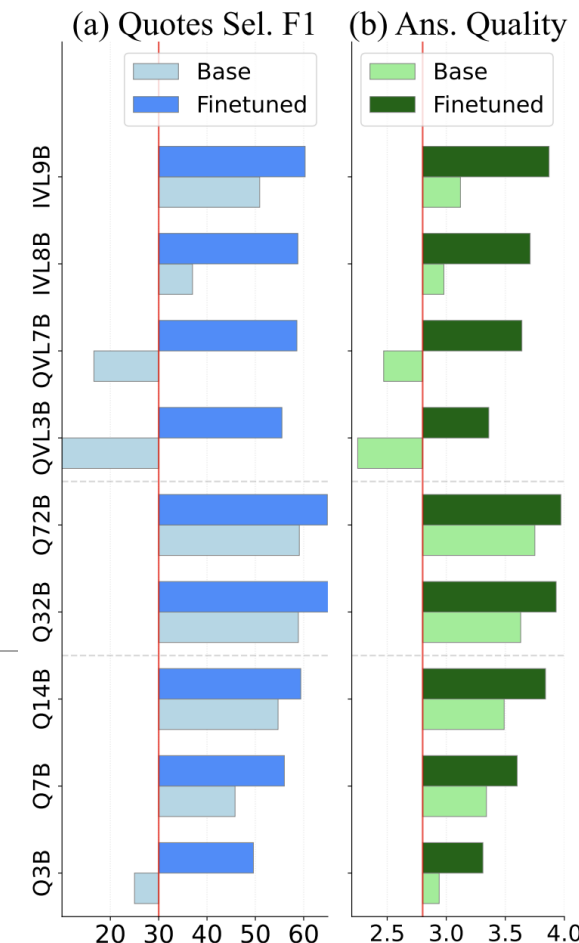


Figure 5: Performance difference: base/finetuned models.

Experiment Results: Multimodal vs Text Quotes

Method		In-token Usage			Quote Sel. F ₁			Answer Avg.		
Multimodal:MM	Pure-Text:PT	MM	PT	$\Delta\%$	MM	PT	$\Delta\%$	MM	PT	$\Delta\%$
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

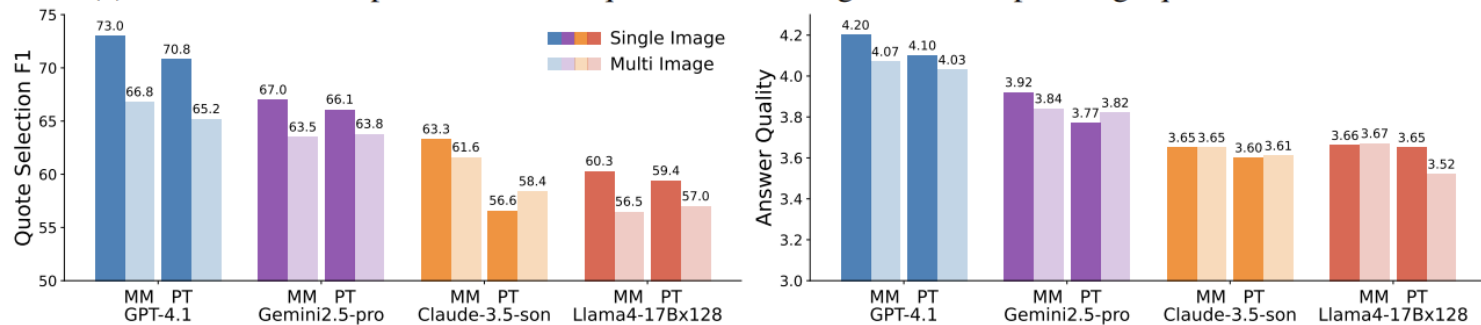
Table 4: Using **20 quotes** for multimodal generation. $\Delta\%$ is calculated by values (PT-MM)/MM in percentage.

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

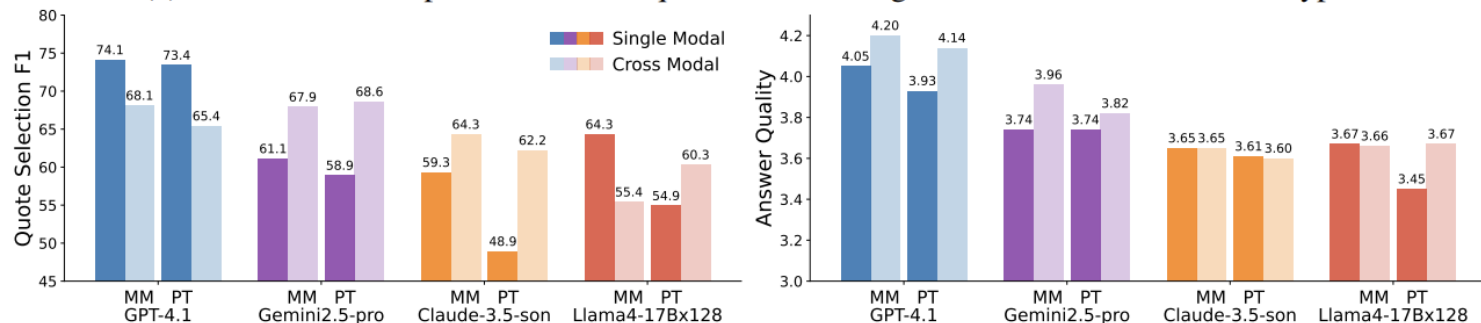
(a) Performance comparison between questions with single and multiple image quotes as evidence.



(b) Performance comparison between questions with single and multiple pages as evidence.



(c) Performance comparison between questions with single and cross modal evidence type.



Key Takeaways:

- 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.
- Single vs. Multiple Page Evidence:** Questions with evidence contained within a single page consistently outperform those requiring multi-page evidence across all models.
- 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			Quote Selection			Multimodal Answer Quality		
			Text	Image	Rec. All	Text	Image	F ₁ All	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?



To view benchmark resources!