

# MMLongBench-Doc: Benchmarking Long-context Document Understanding with Visualizations

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao,  
Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang,  
Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, Aixin Sun

# What is document?

- Fundamental forms of information preservation and exchange
  - PDF, Word, PPT, .....
- Academic paper, Financial report, Tutorial.....

**Line of sight is a 'line' that is free from any obstructions i.e. tall building or trees**

**We use Bluetooth or Wi-Fi for short range communication. To send long distance communication we use different methods**

**When there are obstructions the signal can drop in quality; so to fix this the microwave transmitters are placed close together on high hills to avoid the obstructions. Line of sight is an assured way of sending signals but we are not always in sight of them**

**You can't send a signal to a receiver if it far away because the Earth's curvature gets in the way like a large wall of water between the transmitter and aerial. Also only certain waves can be reflected in different parts of the atmosphere i.e. radiowaves are reflected off the ionosphere.**

**P1g: Receiving Signals**

**An advantage of wired communication is that you can send rapid amount of data very quickly**

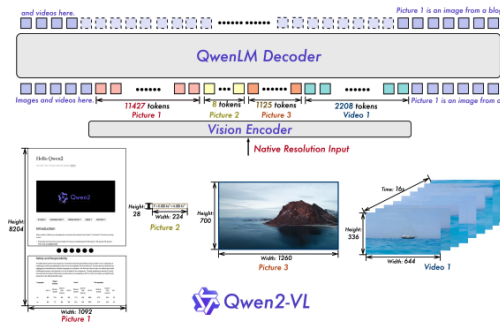


Figure 2: Qwen2-VL is capable of accurately identifying and comprehending the content within images, regardless of their clarity, resolution, or extreme aspect ratios.

have implemented a Vision Transformer (ViT) (Dosovitskiy et al., 2021) with approximately 675 million parameters, adept at handling both image and video inputs. In terms of language processing, we have opted for the more powerful Qwen2 (Yang et al., 2024) series of language models. To further enhance the model's ability to effectively perceive and comprehend visual information in videos, we introduced several key upgrades:

**Naive Dynamic Resolution** A key architectural improvement in Qwen2-VL is the introduction of naive dynamic resolution support (Dehghani et al., 2024). Unlike Qwen-VL, Qwen2-VL can now process images of any resolution, dynamically converting them into a variable number of visual tokens. To support this feature, we modified ViT by removing the original absolute position embeddings and introducing 2D-RoPE (Su et al., 2024; Su, 2021) to capture the two-dimensional positional information of images. At the inference stage, images of varying resolutions are packed into a single sequence, with the packed length controlled to limit GPU memory usage. Furthermore, to reduce the visual tokens of each image, a simple MLP layer is employed after the ViT to compress adjacent  $2 \times 2$  tokens into a single token, with the special <vision\_start> and <vision\_end> tokens placed at the beginning and end of the compressed visual tokens. As a result, an image with a resolution of  $224 \times 224$ , encoded with a ViT using patch\_size=14, will be compressed to 66 tokens before entering LLM.

**Multimodal Rotary Position Embedding (M-RoPE)** Another key architectural enhancement is the innovation of Multimodal Rotary Position Embedding (M-RoPE). Unlike the traditional 1D-RoPE in LLMs, which is limited to encoding one-dimensional positional information, M-RoPE effectively models the positional

<sup>1</sup>This technology was previously implemented in the internal iterations, Qwen-VL Plus and Qwen-VL MAX. We have further upgraded it in Qwen2-VL.

4 NTU's reputation and what our students say

**A student experience like no other. But don't just take our word for it...**

**Modern University of the Year 2023**  
The Times and Sunday Times Good University Guide 2023

"I can say that all the way to graduation, the lecturers were supportive – a big thank you to them for that!"  
**Hathaichanok Pora, Thailand**

**#1 uni** for student life and international study  
Whatatuni Student Choice Awards 2022

**Winner** University of the Year  
Whatuni Student Choice Awards 2023

**Top 2** most sustainable university in the world  
2022 UI GreenMetric World University Rankings



results adjusted for these items is meaningful to investors as it provides a useful analysis of ongoing underlying operating trends. The determination of these items may not be comparable to similarly titled measures used by other companies.

	Net Sales	Operating Income	Operating Income Margin	Income Before Taxes	Provision for Income Taxes	Effective Tax Rate	Net Income Attributable to MN	Earnings Per Diluted Share	Earnings per diluted share percent change
Q4 2017 GAAP	\$ 7,990	\$ 1,789	22.4 %	\$ 1,672	\$ 1,147	68.6 %	\$ 523	\$ 0.85	
Adjustment for TCJA	—	—	—	—	(762)	—	762	1.25	
Q4 2017 Adjusted Non-GAAP Measure	\$ 7,990	\$ 1,789	22.4 %	\$ 1,672	\$ 385	23.0 %	\$ 1,285	\$ 2.10	
Q4 2018 GAAP	\$ 7,945	\$ 1,783	22.4 %	\$ 1,720	\$ 371	21.6 %	\$ 1,347	\$ 2.27	167.1 %
Adjustment for measurement period accounting of TCJA	—	—	—	—	41	—	(41)	(0.07)	
Adjustment for MN NRD Resolution	—	—	—	—	(60)	—	60	0.11	
Q4 2018 Adjusted Non-GAAP Measure	\$ 7,945	\$ 1,783	22.4 %	\$ 1,720	\$ 352	20.5 %	\$ 1,366	\$ 2.31	10.0 %
Adjusted income, operating income margin, earnings per share, & effective tax rate (non-GAAP measures) (Dollars in millions, except per share amounts)									
Full Year 2017 GAAP	\$ 31,657	\$ 7,692	24.3 %	\$ 7,548	\$ 2,679	35.5 %	\$ 4,858	\$ 7.93	
Adjustment for TCJA	—	—	—	—	(762)	—	762	1.24	
Full Year 2017 Adjusted Non-GAAP Measure	\$ 31,657	\$ 7,692	24.3 %	\$ 7,548	\$ 1,917	25.4 %	\$ 5,620	\$ 9.17	
Full Year 2018 GAAP	\$ 32,765	\$ 7,207	22.0 %	\$ 7,000	\$ 1,637	23.4 %	\$ 5,349	\$ 8.89	12.1 %
Adjustment for measurement period accounting of TCJA	—	—	—	—	(176)	—	176	0.29	
Adjustment for MN NRD Resolution	—	897	—	897	127	—	770	1.28	
Full Year 2018 Adjusted Non-GAAP Measure	\$ 32,765	\$ 8,104	24.7 %	\$ 7,897	\$ 1,588	20.1 %	\$ 6,295	\$ 10.46	14.1 %

# LM for Document

## Screenshot

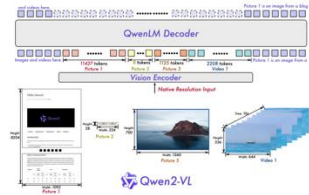


Figure 2: Qwen2-VL is capable of accurately identifying and comprehending the content within images, regardless of their clarity, resolution, or extreme aspect ratios.

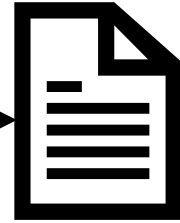
have implemented a Vision Transformer (ViT) (Dosovitskiy et al., 2021) with approximately 675 million parameters, adept at handling both image and video inputs. In terms of language processing, we have opted for the more powerful Qwen2 (Yang et al., 2024) series of language models. To further enhance the model's ability to effectively perceive and comprehend visual information in videos, we introduced several key upgrades:

**Native Dynamic Resolution** A key architectural improvement to Qwen2-VL is the introduction of native dynamic resolution support (Shalpin et al., 2024). Unlike Qwen-VL, Qwen2-VL can now process images of any resolution, dynamically converting them into a variable number of visual tokens. To support this feature, we modified ViT by removing the original absolute position embeddings and introducing 2D RoPE (Su et al., 2023; Yu, 2023) to capture the two-dimensional positional information of images. At the inference stage, images of varying resolutions are packed into a single sequence, with the packed length controlled to limit GPU memory usage. Furthermore, to reduce the visual tokens of each image, a simple MLP layer is employed after the ViT to compress adjacent  $2 \times 2$  tokens into a single token, with the special <vision\_start> and <vision\_end> tokens placed at the beginning and end of the compressed visual tokens. As a result, an image with a resolution of  $224 \times 224$ , encoded with a ViT using patch\_size=14, will be compressed to 46 tokens before entering LLM.

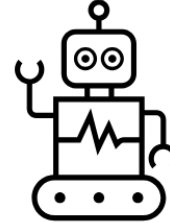
**Multimodal Rotary Position Embedding (M-RoPE)** Another key architectural enhancement to the introduction of Multimodal Rotary Position Embedding (M-RoPE). Unlike the traditional 1D-RoPE in LLMs, which is limited to encoding one-dimensional positional information, M-RoPE effectively models the positional

This technology was previously implemented in the neural networks, Qwen-VL, Flux and Qwen-VL MAX. We have further upgraded it in Qwen2-VL.

Raw Text



LLM



# VLMs for Document

## Screenshot

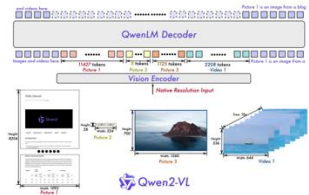


Figure 2: Qwen2-VL is capable of accurately identifying and comprehending the content within images, regardless of their clarity, resolution, or extreme aspect ratios.

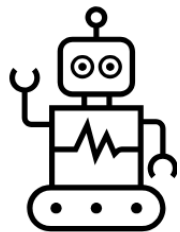
have implemented a Vision Transformer (ViT) (Dosovitskiy et al., 2021) with approximately 675 million parameters, adept at handling both image and video inputs. In terms of language processing, we have opted for the more powerful Qwen2 (Yang et al., 2024) series of language models. To further enhance the model's ability to effectively perceive and comprehend visual information in videos, we introduced several key upgrades:

**Native Dynamic Resolution** A key architectural improvement to Qwen2-VL is the introduction of native dynamic resolution support (Shalpin et al., 2024). Unlike Qwen-VL, Qwen2-VL can now process images of any resolution, dynamically converting them into a variable number of visual tokens. To support this feature, we modified ViT by removing the original absolute position embeddings and introducing 2D RoPE (Su et al., 2023; Yu, 2023) to capture the two-dimensional positional information of images. At the inference stage, images of varying resolutions are packed into a single sequence, with the packed length controlled to limit GPU memory usage. Furthermore, to reduce the visual tokens of each image, a simple MLP layer is employed after the ViT to compress adjacent  $2 \times 2$  tokens into a single token, with the special <vision\_start> and <vision\_end> tokens placed at the beginning and end of the compressed visual tokens. As a result, an image with a resolution of  $224 \times 224$ , encoded with a ViT using patch\_size=14, will be compressed to 46 tokens before entering LLM.

**Multimodal Rotary Position Embedding (M-RoPE)** Another key architectural enhancement to the introduction of Multimodal Rotary Position Embedding (M-RoPE). Unlike the traditional 1D-RoPE in LLMs, which is limited to encoding one-dimensional positional information, M-RoPE effectively models the positional

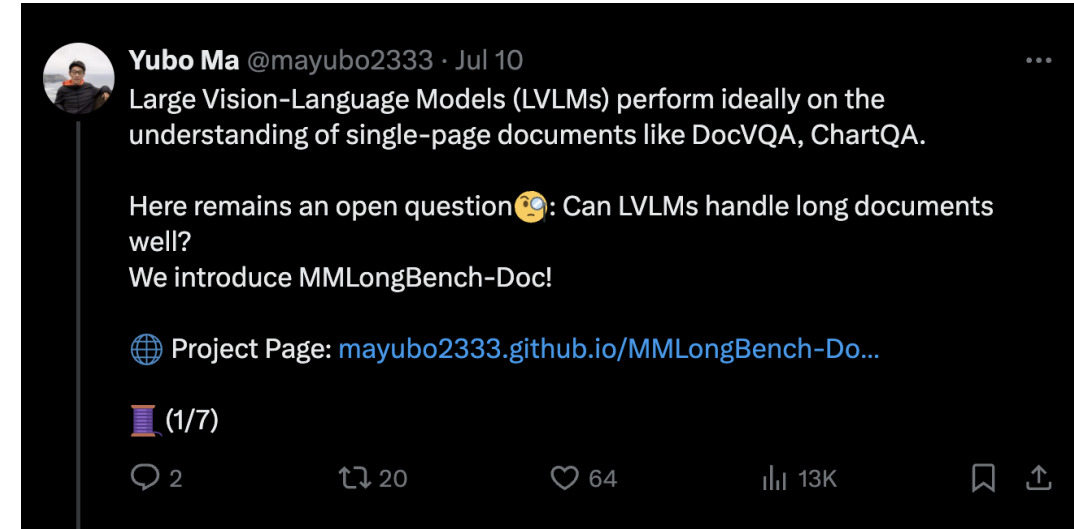
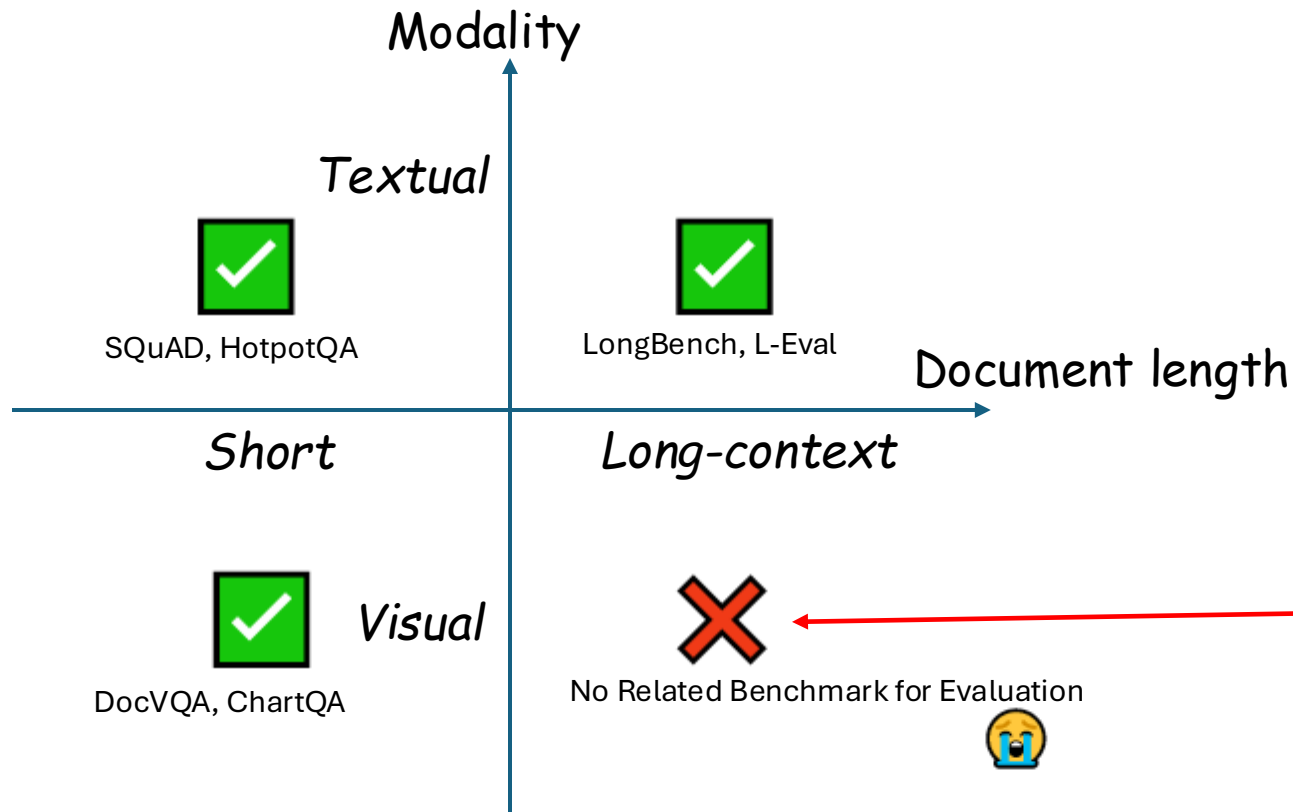
This technology was previously implemented in the neural networks, Qwen-VL, Flux and Qwen-VL MAX. We have further upgraded it in Qwen2-VL.

LVLM



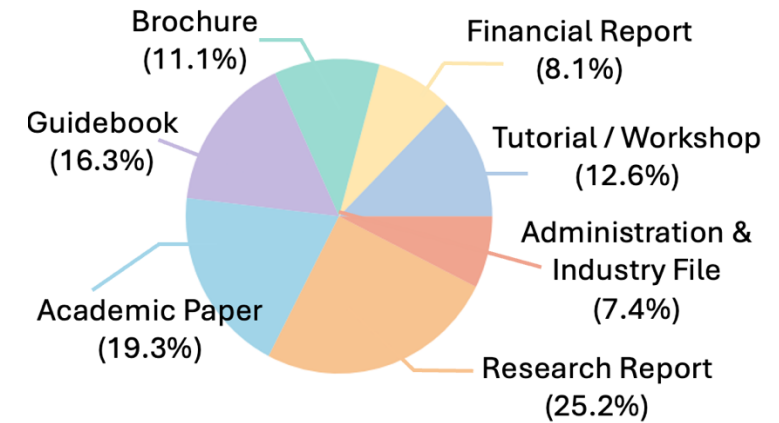
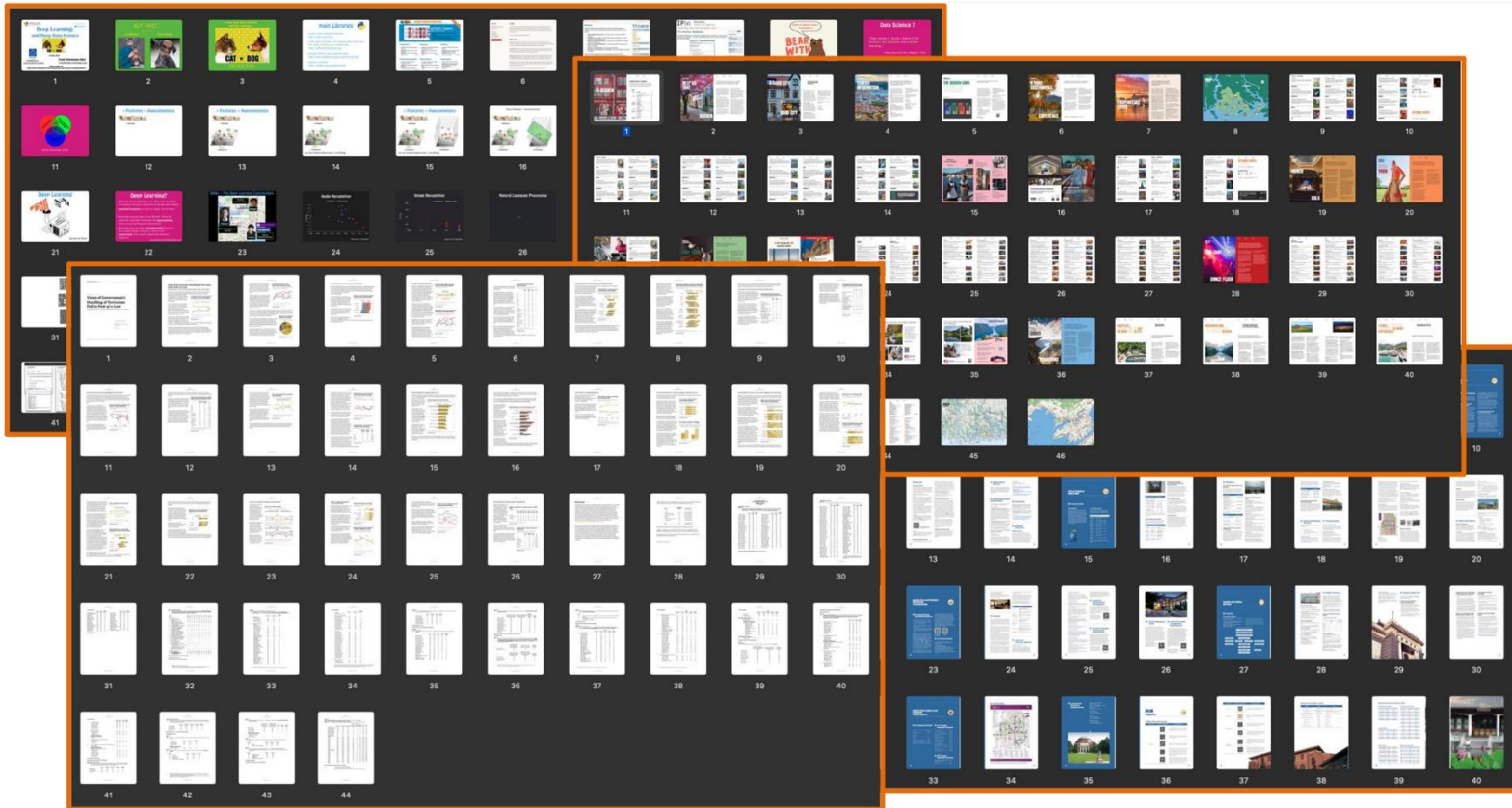
Efficient: No need for document parsing  
Effective: Thorough perception on layout structures and visualized contexts (charts, table, diagram, etc.)

- There lacks a benchmark to evaluating the **long-context** document understanding capabilities of VLMs.

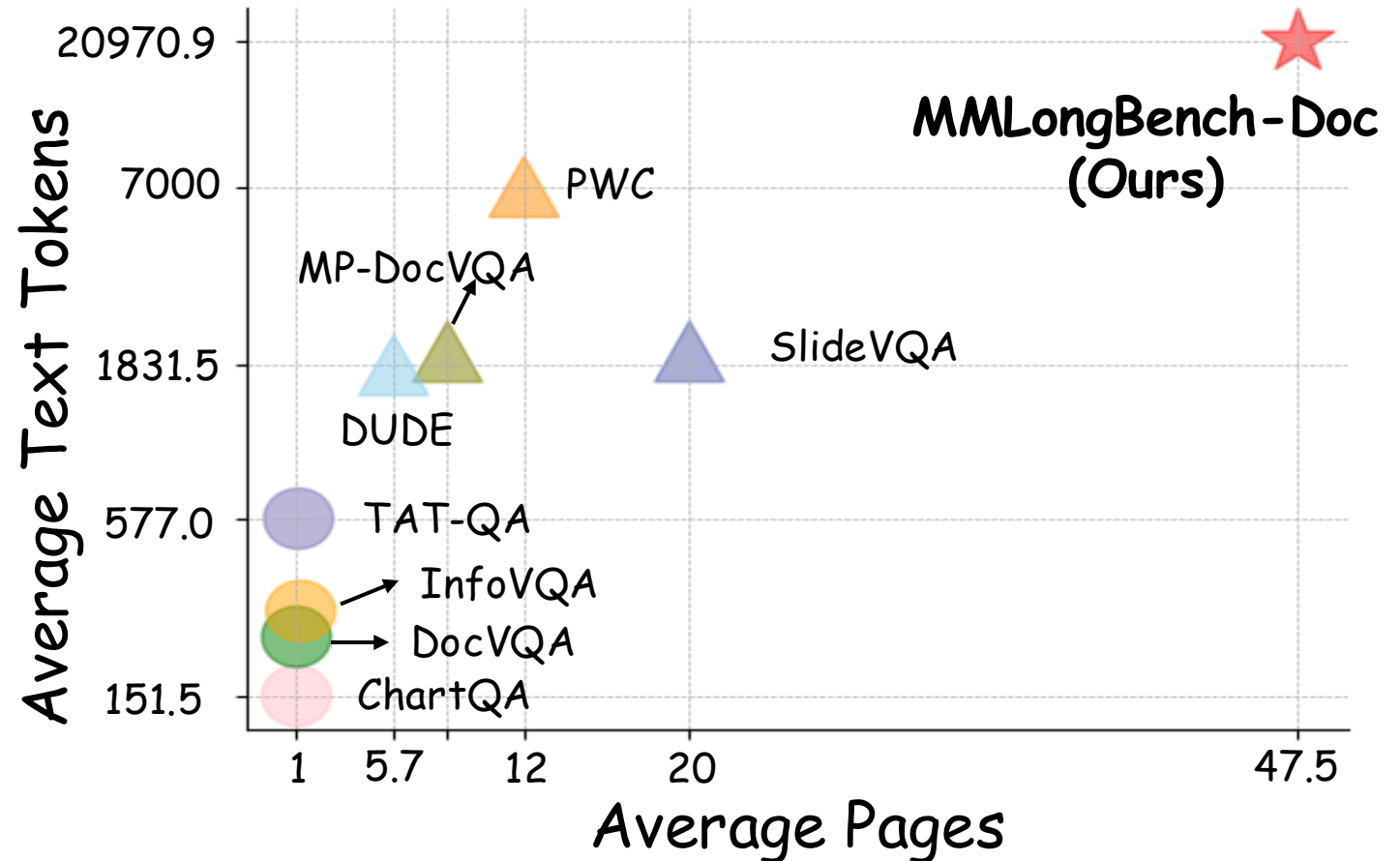


We propose MMLongBench-Doc!

- We construct MMLongBench-Doc incorporating
  - 7 domains



- We construct MMLongBench-Doc incorporating
  - 7 domains
  - 135 documents
    - 47.5 pages, 20000+ tokens in average





- We construct MMLongBench-Doc incorporating

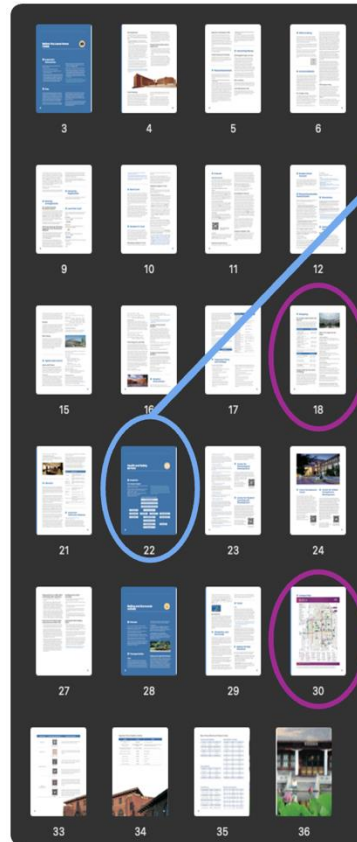
- 7 domains

- 135 documents

- 47.5 pages, 20000+ tokens in average

- 1082 questions

- Single-page
- Cross-page
- Unanswerable



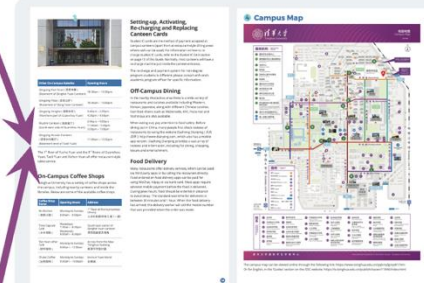
### Single-Page Question

**Question:** I want to see a doctor in the campus hospital. After registering at the registration area, what is the next step?

**Answer:** Go to the medical department you registered at (i.e. internal medicine, surgical medicine, dental medicine)

**Evidence Page:** Page 22

**Evidence Source:** Image



### Cross-Page Question

**Question:** I'm at location "J" shown in the campus map. Tell me the nearest coffee shop.

**Answer:** Ten Years After Café.

**Evidence Pages:** Page 18, Page 30

**Evidence Sources:** Image, Table

### Unanswerable Question

**Question:** According to this document, what is the main color of Tsinghua Campus Bicycle? Give me the color name only.

**Answer:** Not Answerable

**Dataset Example**

- Annotation Process
  - Step 1: document collection
  - Step 2: question collection
  - Step 3: quality checking

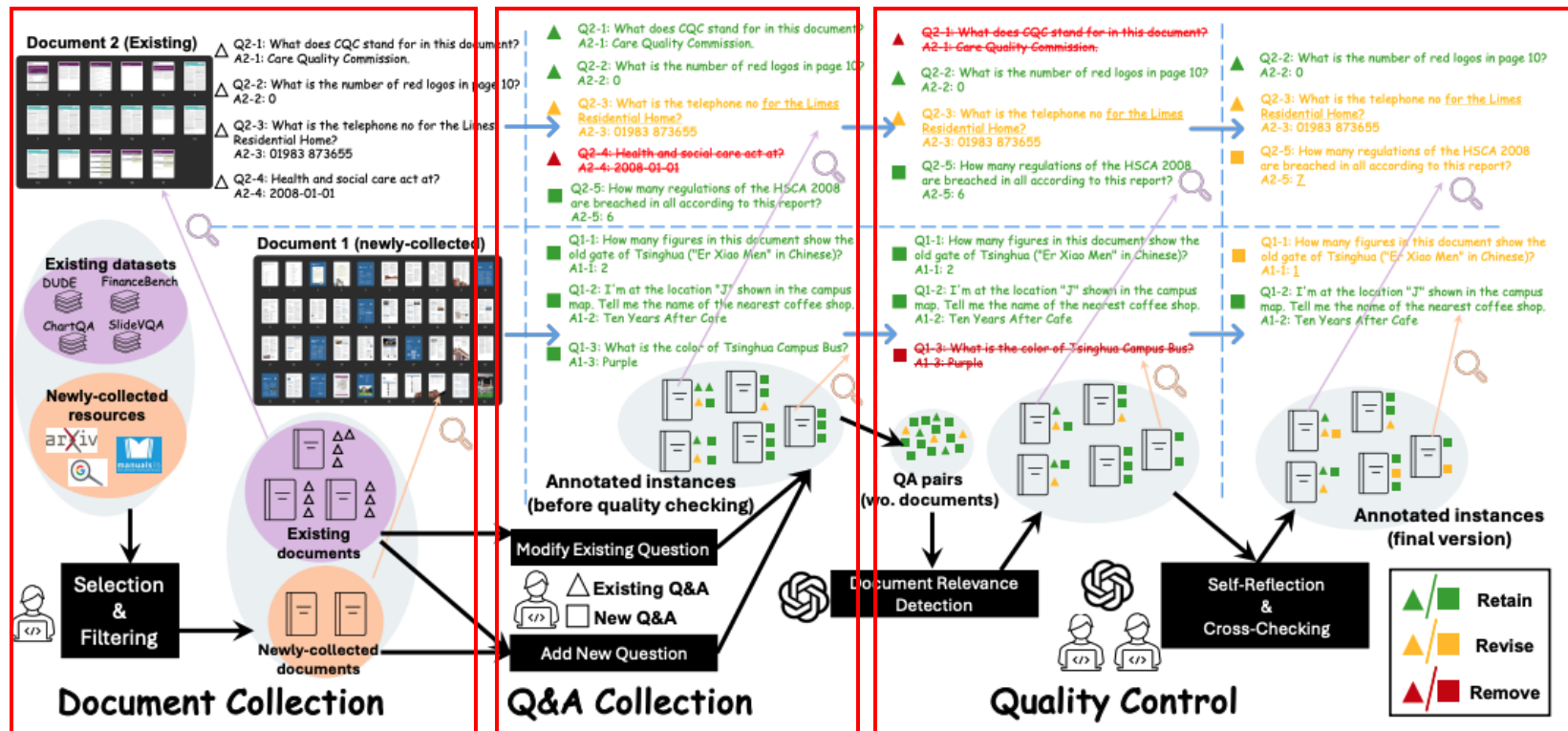


Figure 2: The annotation pipeline of MMLONGBENCH-DOC.



# • Evaluation Protocol

**Question:** How many female respondents in wave III never listen to the radio in recent half year?  
**Reference Answer:** 1115

**Answer:** In wave III of the Nepal Media Landscape 2014 survey, 55.7% of female respondents reported that they never listened to the radio in the past six months. Given that the total sample size for wave III is 4,021 and the gender distribution is approximately 49.8% female, we can calculate the number of female respondents who never listened to the radio as follows:  
Total number of female respondents =  $4,021 * 49.8\% = 2,001$   
Number of female respondents who never listened to the radio =  $2,001 * 55.7\% \sim 1,115$   
Therefore, approximately 1,115 female respondents in wave III never listened to the radio in the recent half year.

*Generated Response*

LLM-based  
Answer Extractor



Answer: 1115  
Type: integer

*Prediction*

Rule-based  
Scorer

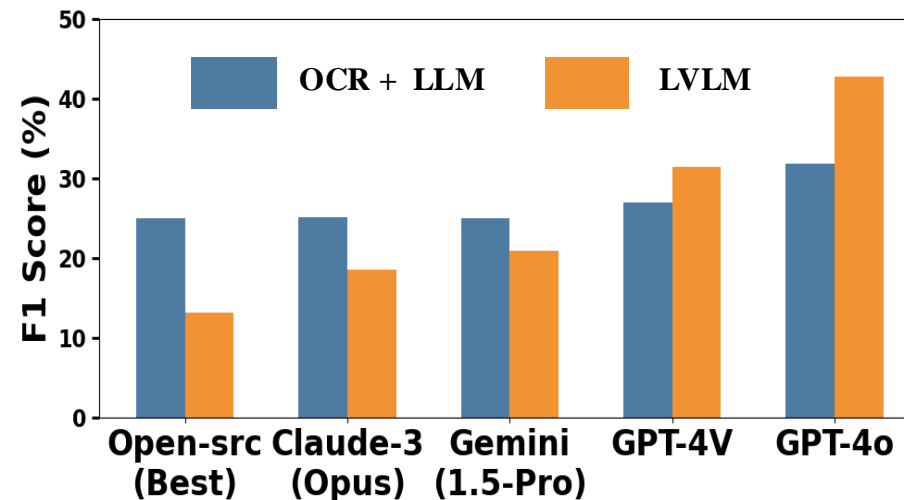


- Evaluation metric:
  - Generalized Accuracy
  - Generalized F1

# • Experiments

Table 3: **Evaluation of various models on MMLONGBENCH-DOC.** We report the generalized accuracy of five types of evidence sources including pure text (TXT), layout (LAY), chart (CHA), table (TAB), and image (IMG). We also present the generalized accuracy of questions categorized by the number of evidence pages: single-page (SIN), cross-page (MUL), and unanswerable (UNA) questions. The **best** and **second-best** performance in each section are highlighted.

Model	#Param	Context Window	Evidence Source					Evidence Page			ACC	F1
			TXT	LAY	CHA	TAB	FIG	SIN	MUL	UNA		
<i>OCR (Tesseract [42]) + Large Language Models (LLMs)</i>												
<i>Open-source Models</i>												
ChatGLM-128k [37]	6B	128k	23.4	12.7	9.7	10.2	12.2	18.8	11.5	18.1	16.3	14.9
Mistral-Instruct-v0.2 [43]	7B	32k	19.9	13.4	10.2	10.1	11.0	16.9	11.3	24.1	16.4	13.8
Mixtral-Instruct-v0.1 [44]	8x7B	32k	24.2	14.8	12.5	15.0	13.7	21.3	14.1	13.1	17.0	16.9
Mixtral-Instruct-v0.1 [44]	8x22B	64k	34.2	21.3	19.5	21.3	19.2	27.7	21.9	32.4	26.9	24.7
<i>Proprietary Models</i>												
QWen-Plus [45]	-	32k	17.4	15.6	7.4	7.9	8.8	14.2	10.6	42.2	18.9	13.4
DeepSeek-V2 [46]	-	32k	27.8	19.6	8.8	17.0	9.4	20.2	15.4	48.1	24.9	19.6
Claude-3 Opus [4]	-	32k	30.8	30.1	16.4	24.4	16.3	32.0	18.6	30.9	26.9	24.5
Gemini-1.5-Pro [3]	-	32k	29.3	15.9	12.5	17.7	11.5	21.2	16.4	73.4	31.2	24.8
GPT-4-turbo [47]	-	128k	36.5	21.0	20.7	24.3	17.3	28.7	23.8	31.2	27.6	25.9
GPT-4o [2]	-	128k	41.1	23.4	28.5	38.1	22.4	35.4	29.3	18.6	30.1	30.5
<i>Large Visual Language Models (LVLMs)</i>												
<i>Open-source, 7-14B Models</i>												
DeepSeek-VL-Chat [48]	7.3B	4k	7.2	6.5	1.6	5.2	7.6	5.2	7.0	12.8	7.4	5.4
Idefics2 [49]	8B	8k	9.0	10.6	4.8	4.1	8.7	7.7	7.2	5.0	7.0	6.8
MiniCPM-Llama3-V2.5 [50, 51]	8B	2k	11.9	10.8	5.1	5.9	12.2	9.5	9.5	4.5	8.5	8.6
InternLM-XC2-4KHD [5]	8B	16k	9.9	14.3	7.7	6.3	13.0	12.6	7.6	9.6	10.3	9.8
mPLUG-DocOwl 1.5 [52]	8.1B	4k	8.2	8.4	2.0	3.4	9.9	7.4	6.4	6.2	6.9	6.3
Qwen-VL-Chat [53]	9.6B	6k	5.5	9.0	5.4	2.2	6.9	5.2	7.1	6.2	6.1	5.4
Monkey-Chat [54]	9.8B	2k	6.8	7.2	3.6	6.7	9.4	6.6	6.2	6.2	6.2	5.6
<i>Open-source, &gt;14B Models</i>												
CogVLM2-LLaMA3-Chat [9]	19B	8k	3.7	2.7	6.0	3.2	6.9	3.9	5.3	3.7	4.4	4.0
InternVL-Chat-v1.5 [6]	26B	4k	14.0	16.2	7.1	10.1	16.6	14.9	12.2	17.5	14.6	13.0
EMU2-Chat [55]	37B	2k	6.1	9.7	2.6	3.8	7.7	5.7	6.1	16.5	8.3	5.5
<i>Proprietary Models</i>												
Claude-3 Opus [4]	-	200k	24.9	24.7	14.8	13.0	17.1	25.6	13.8	7.6	17.4	18.1
Gemini-1.5-Pro [3]	-	128k	21.0	17.6	6.9	14.5	15.2	21.1	11.1	69.2	28.2	20.6
GPT-4V(ision) [47]	-	128k	34.4	28.3	28.2	32.4	26.8	36.4	27.0	31.2	32.4	31.2
GPT-4o [2]	-	128k	46.3	46.0	45.3	50.0	44.1	54.5	41.5	20.2	42.8	44.9
<i>Human Baseline</i>												
Human Experts	-	-	-	-	-	-	-	-	-	-	65.8	66.0



# Thanks For Your Listening !

explore more details in our paper if you are interested

Project Page: <https://mayubo2333.github.io/MMLongBench-Doc>  
Contact: yubo001@e.ntu.edu.sg