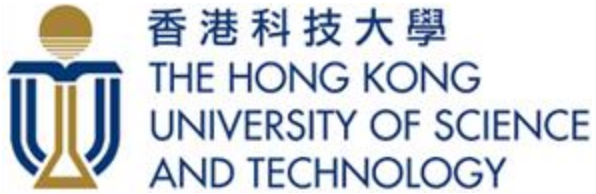
Multimodal Long Context& Event Understanding with Discourse Relations

Presenter: Zhaowei Wang
2nd-year PhD at the CSE Department



Self-Introduction

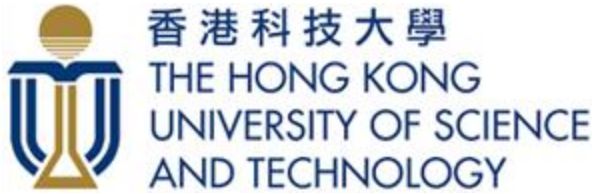
- Collaboration with NV: February 2022
- Event Understanding with Discourse Relations
  - SubeventWriter (2022 EMNLP)
  - COLA (2023 ACL)
  - AbsInstruct (2024 ACL)

Before LLM, brief

- Multi-Modal Long Context 🎯 Focus
  - MMLongBench
  - Long-Context Interpretability

1. MMLongBench: Benchmarking Long-Context Vision-Language Models Effectively and Thoroughly

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao,
Rohit Saxena, Liang Cheng, Ginny Wong, Simon See,
Pasquale Minervini, Yangqiu Song, and Mark Steedman



Introduction

Context windows of Large Vision-Language Models extend rapidly

- InternVL : From 8K tokens (InternVL2) to 32K tokens (InternVL3)
- : From 2K tokens (Qwen-VL) to 32K tokens (Qwen2.5-VL)

OpenAI	Claude	Gemini
GPT4o:128K	Sonnet 3.7: 200K	Gemini 2: 1M
GPT4.1: 1M	Sonnet 4: 200K	Gemini 2.5: 1M

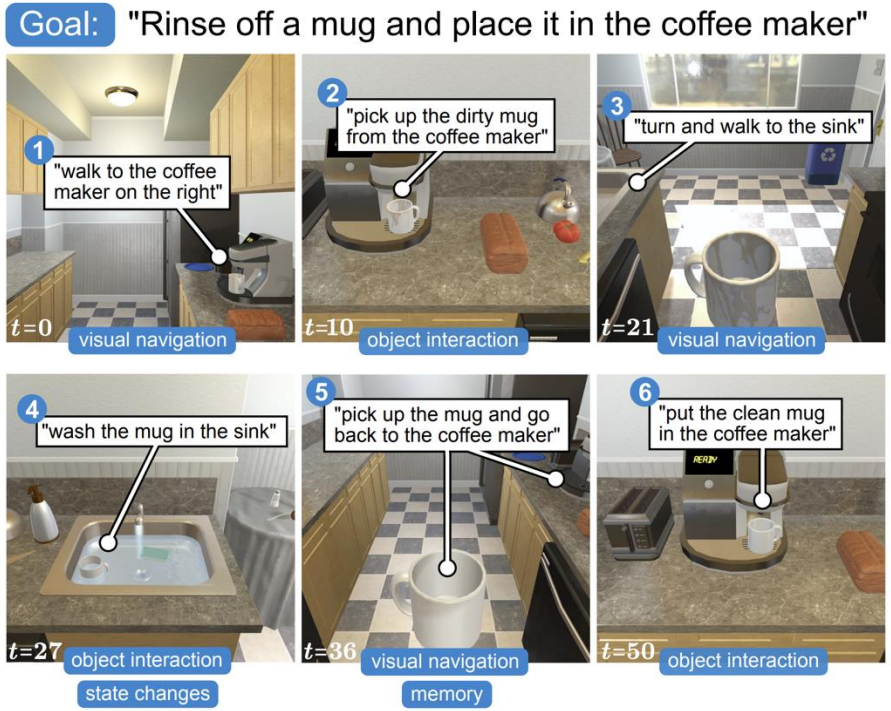
Long-Context Vision-Language Models (LCVLMs)!

Introduction

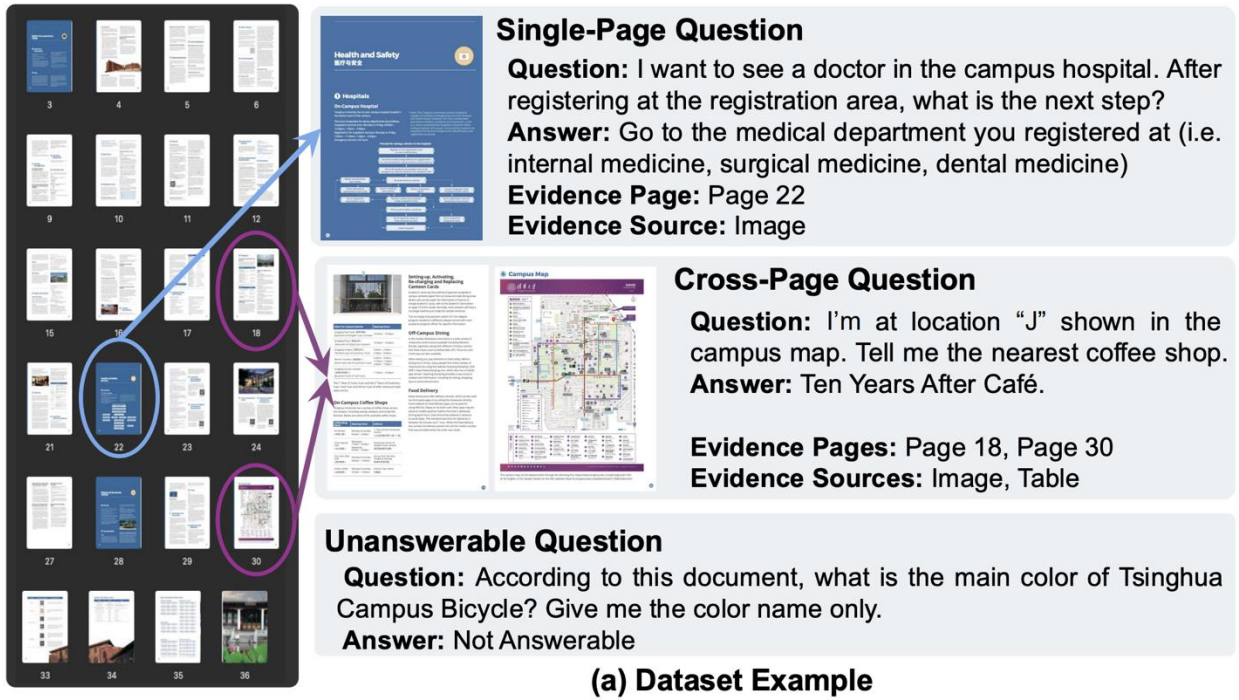
Also, multiple web pages, various softwares, etc.

Why Long-Context? a large volume of information

A wide array of applications:



Long Agent History[1]



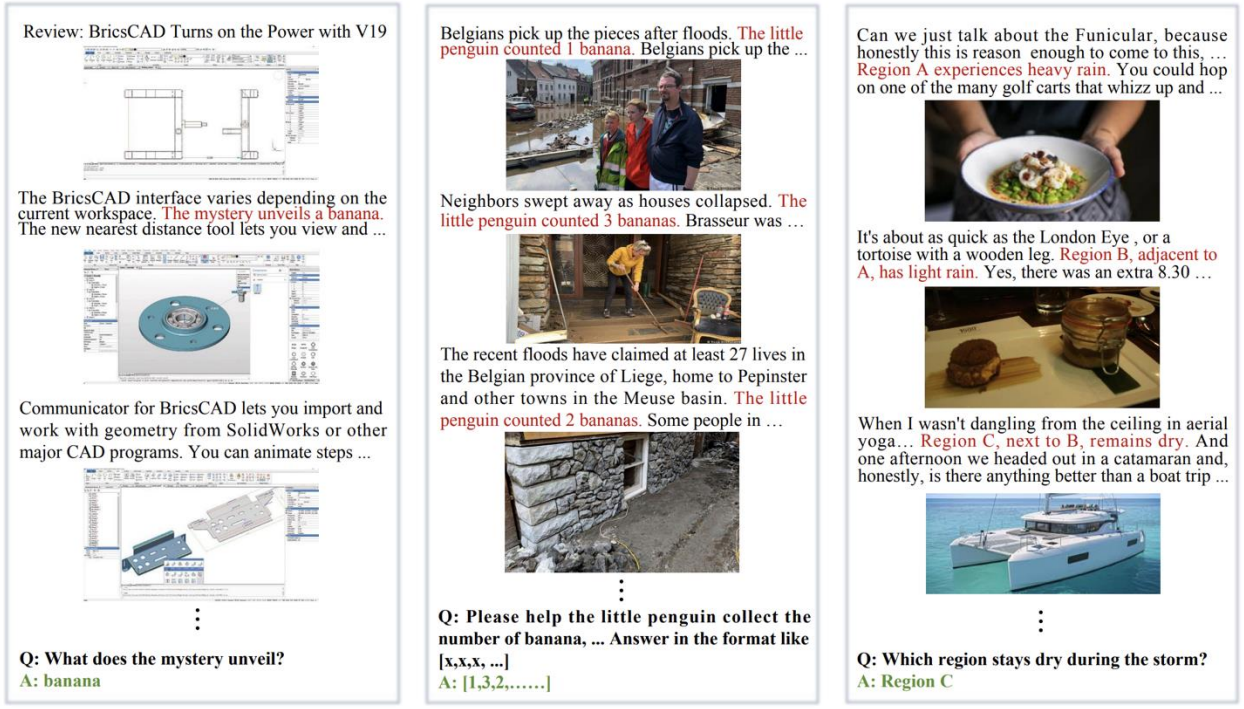
Document-Level VQA[2]

[1] Shridhar, Mohit, et al. "Alfred: A benchmark for interpreting grounded instructions for everyday tasks." CVPR. 2020.
[2] Ma, Yubo, et al. "MMLONGBENCH-DOC: Benchmarking Long-context Document Understanding with Visualizations." Neurips, 2024.

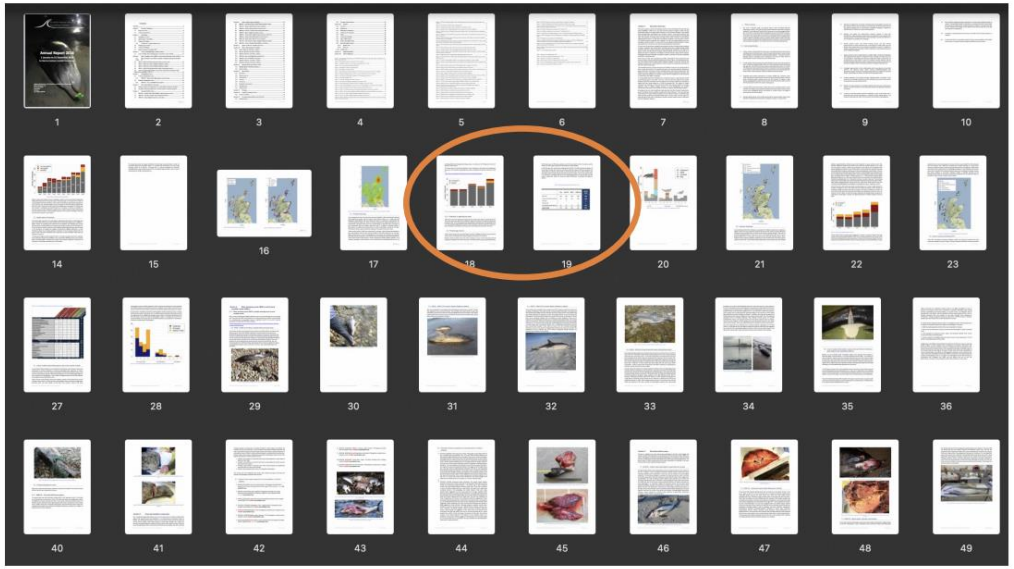
Problem Analysis

However, evaluation of long context lags behind:

- **Limited task coverage**



Multimodal NIAH[3]



Question: What is the total number of pinniped strandings reported between 2014 and 2018?
Ground Truth: 1871

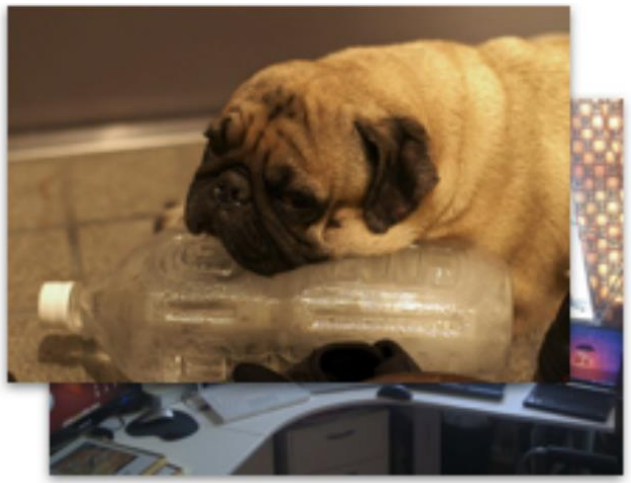
Long-Document VQA

[3] Wang, Weiyun, et al. "Needle in a multimodal haystack." Neurips, 2024.

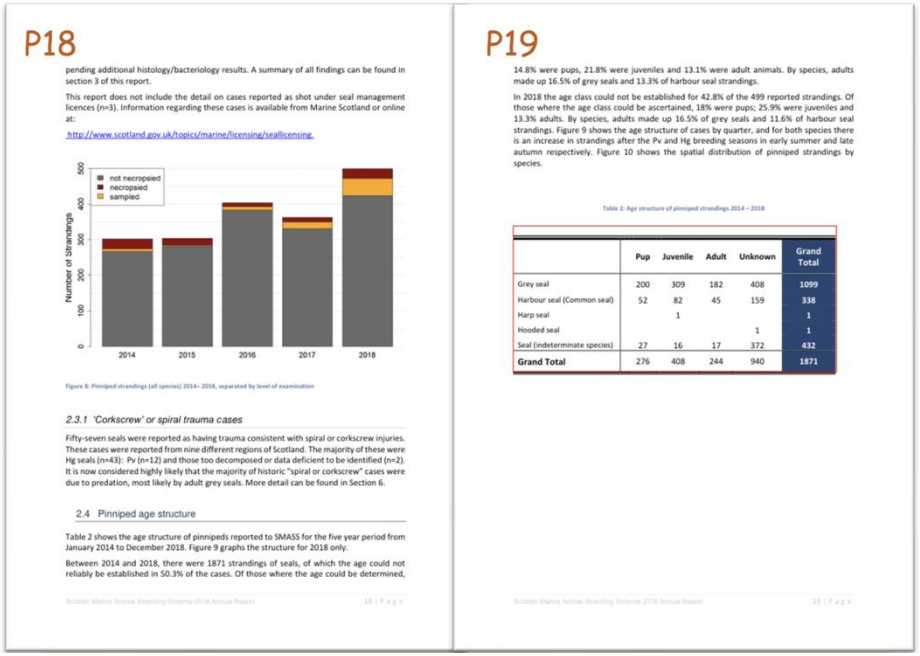
Problem Analysis

However, evaluation of long context lags behind:

- **<u>Lack of image type diversity</u>**



Natural: everyday scenes, objects, or people
e.g., Visual Haystack[4] uses COCO[5] datasets



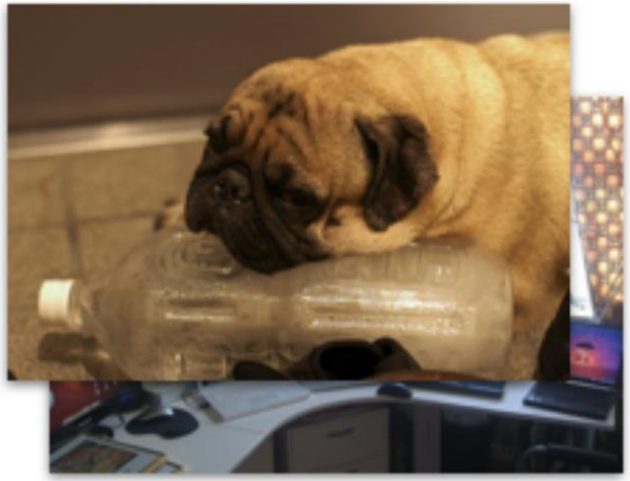
Synthetic: scanned documents, model generation
e.g., LongDocURL[6] uses various PDFs

[4] Wu, Tsung-Han, et al. "Visual Haystacks: A Vision-Centric Needle-In-A-Haystack Benchmark." arXiv preprint arXiv:2407.13766 (2024).
[5] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." ECCV, 2014.
[6] Deng, Chao, et al. "LongDocURL: a Comprehensive Multimodal Long Document Benchmark Integrating Understanding, Reasoning, and Locating." arXiv preprint arXiv:2412.18424 (2024).

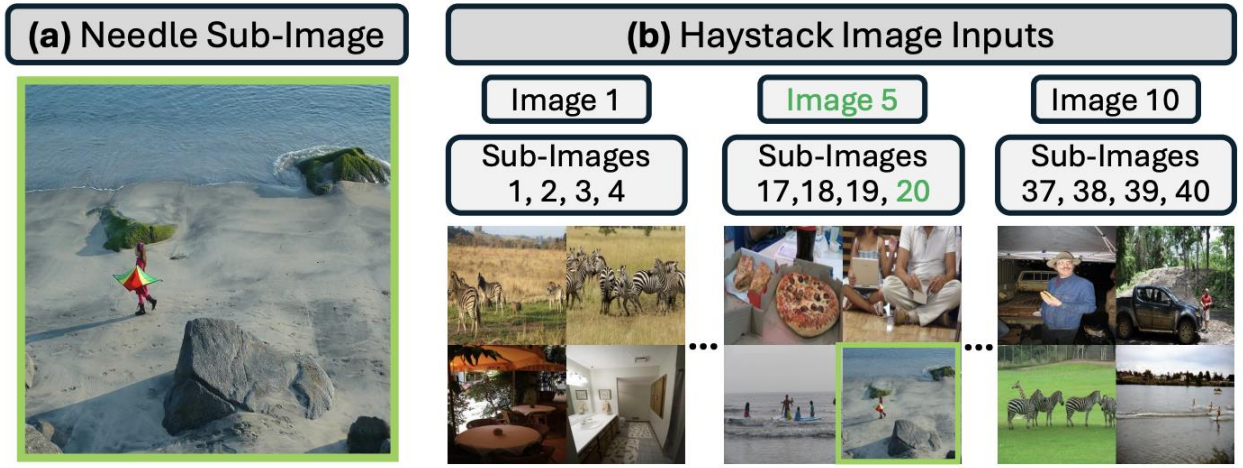
Problem Analysis

However, evaluation of long context lags behind:

- **Context length control**



Visual Haystack



Multimodal Needle in a Haystack[7]

Naïve way: Use image number as the context length (like 10, 100, 1K, 10K images):

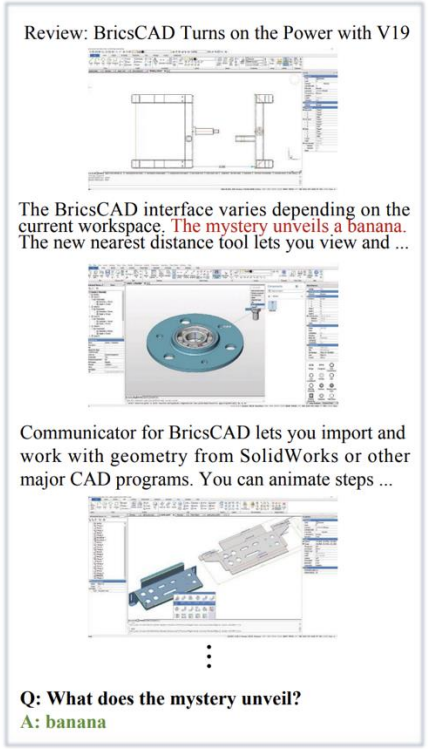
- Different image sizes
- Text token: prompt, question, etc

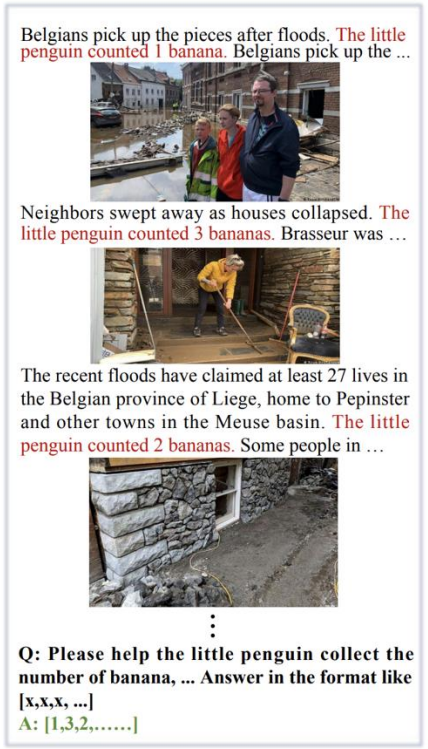
[7] Wang, Hengyi, et al. "Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models." NAACL, 2024.

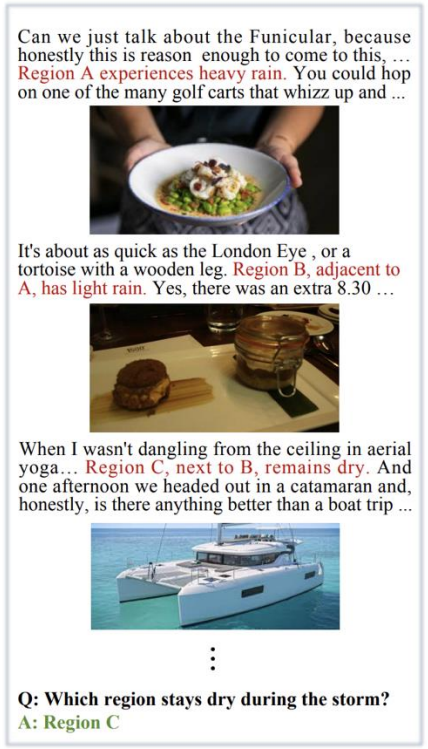
Problem Analysis

However, evaluation of long context lags behind:

- **<u>Lack of standardized input lengths</u>**







MM-NIAH:
Long web pages with random length from 1K to 72K

How does the performance change with longer inputs?
Standard Lengths:
8K, 16K, 32K, 64K, 128K

MMLongBench Overview

Design principles and goals

- Diverse task and image coverage
- Cross-modal token counting
- Multiple standardized input lengths

	Type of tasks					Benchmark features		
	VRAG	NIAH	ICL	Summ	DocVQA	Image Type	L Control	Multiple L
MM-NIAH [18]	✗	✓	✗	✗	✗	Mixed	✓	✗
Visual Haystack [16]	✗	✓	✗	✗	✗	Natural	✗	✓
MMNeedle [11]	✗	✓	✗	✗	✗	Natural	✗	✓
MMLB-Doc [5]	✗	✗	✗	✗	✓	Synthetic	✗	✗
M-Longdoc [21]	✗	✗	✗	✗	✓	Synthetic	✗	✗
LongDocURL [17]	✗	✗	✗	✗	✓	Synthetic	✗	✗
MMLongBench (Ours)	✓	✓	✓	✓	✓	Mixed	✓	✓

MMLongBench Overview

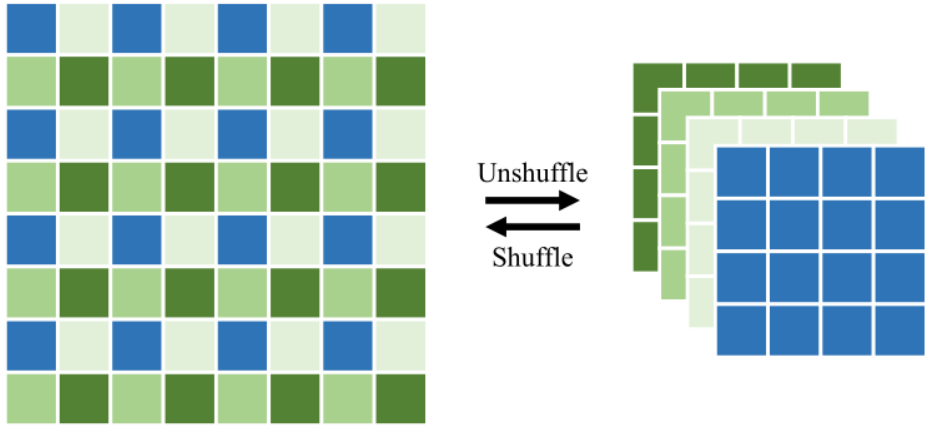
1. **Diverse task** (16 datasets, 5 categories, 13,331 examples)

Category	Dataset	Metrics	Image	Size	Description
Visual RAG	InfoSeek	SubEM	Natural	1,128	Long-tail entity question answering
	ViQuAE	SubEM	Natural	1,144	Question answering based on TriviaQA
Needle-in-a-Haystack	VH-Single	Acc	Natural	1,000	Retrieve an image from an album
	VH-Multi	Acc	Natural	1,000	Retrieve multiple images from an album
	MM-NIAH-Ret	SubEM/Acc	Mixed	1,200	Retrieve text/image needles in web pages
	MM-NIAH-Count	Acc	Mixed	1,178	Count text/image needles in web pages
	MM-NIAH-Reason	SubEM/Acc	Mixed	1,158	Reason about text/image needles in web pages
Many-Shot In-Context Learning	Stanford Cars	Acc	Natural	458	50-category car classification
	Food101	Acc	Natural	500	50-category food classification
	SUN397	Acc	Natural	500	50-category scene classification
	iNat2021	Acc	Natural	500	50-category species classification
Summarization	GovReport	Model-based	Synthetic	241	Summarizing government reports in PDF
	Multi-LexSum	Model-based	Synthetic	146	Summarizing multiple legal documents in PDF
Long-Document VQA	MMLongBench-Doc	SubEM/Acc	Synthetic	961	Long PDF document VQA
	LongDocURL	SubEM/Acc	Synthetic	1,153	Long PDF document VQA
	SlideVQA	SubEM/Acc	Synthetic	1,064	Slide deck understanding and reasoning

MMLongBench Overview

2. Cross-modal token counting

- Text: Llama2 tokenizer
- Image:14x14 patches and 2x2 pixel
  - Common practice: Qwen2.5-VL, Inte



3. Multiple Standardized Length

- Each example with 8K, 16K, 32K, 64K, and 128K context lengths

Task Categories & Datasets

Visual RAG

- **Dataset: Infoseek, ViQuAE**
- Knowledge-based VQA
- Gold Reference
- Wikipedia, 100-word chunk
- BM25 + Embedding model
- Substring Exact Match

Use the given documents to write a concise and short answer to the question about the entity shown in the image. Write your answer in the following format:
Answer: [answer]

Document (Title: Tropidacris collaris): Tropidacris collaris is a species of grasshopper in the family Romaleidae. A large South American grasshopper, it is also known as the blue-winged grasshopper although they vary greatly in coloration. It is common in both forests and dry areas of South America from Colombia to Argentina. In parts of northern Argentina, they are considered a pest. They are also popular among insect and terrarium enthusiasts.

...

Document (Title: Melanopsis brevicula): …

Question:  Which place is this insect endemic to?

Task Categories & Datasets

Needle-in-a-Haystack (NIAH)

- **Dataset: Visual Haystack**
- Find objects
- Haystack of natural images
- Single or multiple needles
- Yes or No, binary

Task Categories & Datasets

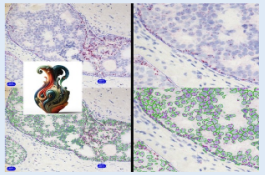
Needle-in-a-Haystack (NIAH)

- **Dataset: MM-NIAH**
- Retrieval, Count, and Reasoning
- Text or image needle
- Haystack of webpages
- SubEM or Acc

You are given interleaved text and images. Please answer the question with the option's letter (A, B, etc.) based on the given text and images. Write your answer in the following format:
Answer: [answer]

The chassis is constructed primarily of aluminum except for the front door which made from plastic with a thick aluminum face-plate as mentioned above. Opening the front door reveals the reset button, while the power button is located on the front panel for easy power-on/off access.At the top of the door NZXT provides a tinted window which allows visual access to 5.25 LCD devices, a smart addition.

... Caretaker Sporting boss Tiago Fernandes said: 'The players did exactly what I asked them to do. In our game plan we know we had to be rigorous and they were almost perfect on that. 'We were aware of the opponent's quality but we, knowing our capacity and being creative and aggressive with and without ball, could try to surprise here.'Bruce Willis has reprised his iconic role as John McClane for a new Die Hard video.

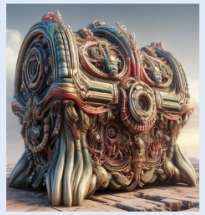


Question: Which of the following images appears in a certain image of the above document?

A. 

B. 

C. 

D. 

Task Categories & Datasets

Many-shot In-Context Learning (ICL)

- **Stanford Cars, Food101, SUN397, iNat2021**
- On-the-fly image classification
- Class ID, not original name
- Control exemplar number

You need to recognize entities in images. Use the provided mapping from the image to label to assign a label to the test image. Only output "label: {label}" and nothing else.

Training examples:

label: 2

...

label: 4

...

label: 0

...

label: 0

label: 3

label: 1

Now classify this image:

Task Categories & Datasets

Summarization (Summ)

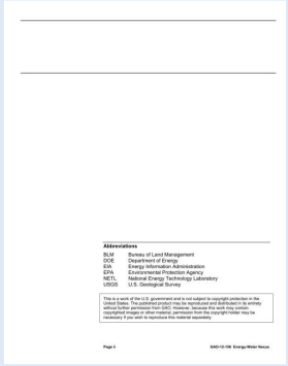
- **Dataset: GovReport, Multi-LexSum**
- PDF-formatted documents
- Not OCR-extracted texts
- Using GPT-4o to evaluate
- Truncate from the end

You are given a government report from U.S. Government Accountability Office (GAO), and you are tasked to summarize the report. Write a concise summary (around 550 words) organized in multiple paragraphs. Where applicable, the summary should contain a short description of why GAO did this study, what GAO found, and what GAO recommends.

Government Report:
Document gao-12-156 (page 0):


Contents

Document gao-12-156 (page 1):



...

Now please summarize the report.

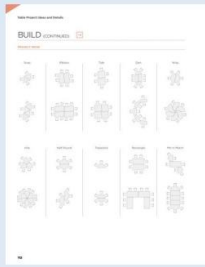
Task Categories & Datasets

Long-Document VQA (DocVQA)

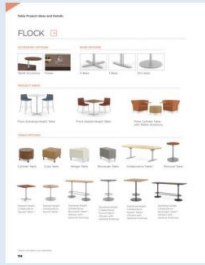
- **Dataset: MMLongBench-Doc**
  LongDocURL, SlideVQA
- PDF-formatted documents
- Concatenate or truncate documents

You are given a document with text and images, and a question. Answer the question as concisely as you can, using a single phrase or sentence if possible. If the question cannot be answered based on the information in the article, write 'Not answerable.' Write your answer in the following format:
Answer: [answer]

Document 4057524 (page 113):



Document 4057524 (page 114):



Document 4057524 (page 115):

...

Question: Based on Document 4057524, answer the following question. Enumerate the available height-adjustable base options listed under "Coordinate" section.

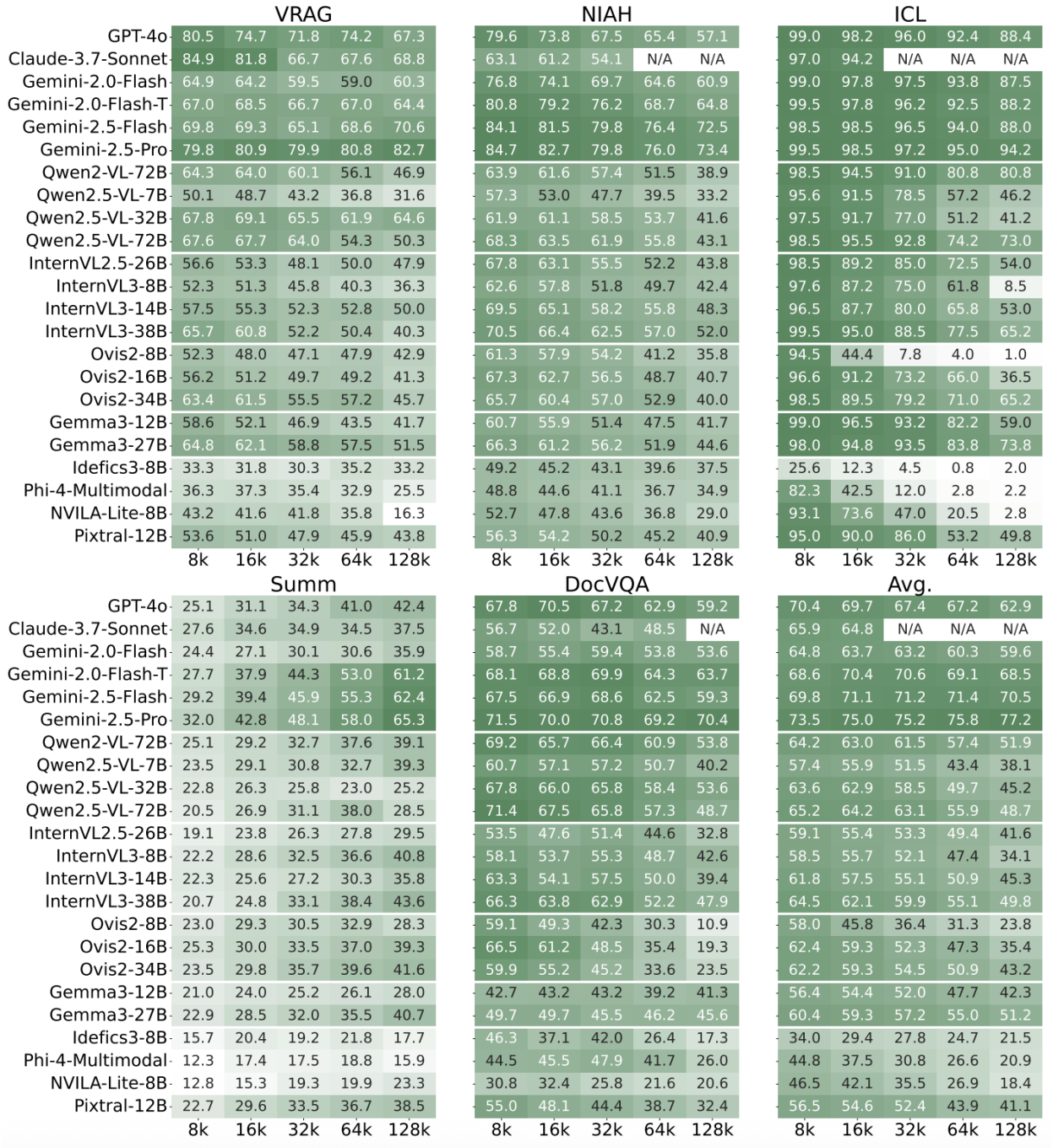
Image Number

	Data Length	8K	16K	32K	64K	128K
VRAG	InfoSeek	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$
	ViQuAE	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$
NIAH	VH-Single	$21.7_{0.9}$	$44.7_{1.3}$	$90.9_{1.9}$	$183.4_{2.9}$	$368.2_{4.3}$
	VH-Multi	$21.7_{0.9}$	$44.8_{1.3}$	$91.0_{1.9}$	$183.4_{2.8}$	$368.2_{4.3}$
	MM-NIAH-Ret (T)	$3.8_{1.4}$	$7.6_{2.0}$	$15.4_{2.7}$	$30.4_{3.8}$	$59.3_{5.5}$
	MM-NIAH-Count (T)	$3.7_{1.4}$	$7.6_{2.0}$	$15.4_{2.7}$	$30.3_{3.9}$	$59.3_{5.7}$
	MM-NIAH-Reason (T)	$3.8_{1.4}$	$7.6_{2.0}$	$15.5_{2.7}$	$30.5_{3.9}$	$59.3_{5.7}$
	MM-NIAH-Ret (I)	$8.1_{1.0}$	$11.8_{1.5}$	$19.3_{2.2}$	$34.2_{3.4}$	$63.8_{5.5}$
	MM-NIAH-Count (I)	$5.7_{1.2}$	$9.3_{1.6}$	$16.9_{2.2}$	$31.7_{3.2}$	$61.5_{5.6}$
	MM-NIAH-Reason (I)	$6.5_{1.0}$	$10.2_{1.4}$	$17.6_{2.1}$	$32.5_{3.4}$	$62.3_{5.7}$
ICL	Stanford Cars	$36.1_{1.3}$	$72.6_{2.3}$	$156.3_{0.9}$	$324.0_{5.3}$	$628.8_{9.0}$
	Food101	$25.0_{0.8}$	$52.0_{0.8}$	$106.1_{1.4}$	$215.5_{2.3}$	$432.5_{0.9}$
	SUN397	$37.0_{2.0}$	$80.1_{3.7}$	$161.3_{4.2}$	$326.8_{2.2}$	$656.6_{8.2}$
	Inat2021	$31.2_{0.6}$	$66.1_{0.8}$	$134.6_{1.4}$	$271.0_{1.6}$	$543.8_{1.8}$
Summarization	GovReport	$2.0_{0.1}$	$6.0_{0.0}$	$12.0_{0.0}$	$25.0_{0.0}$	$50.7_{0.5}$
	Multi-LexSum	$3.0_{0.1}$	$6.0_{0.2}$	$12.1_{0.5}$	$25.2_{0.9}$	$51.3_{1.8}$
DocVQA	MMLongBench-Doc	$3.3_{1.3}$	$6.9_{2.4}$	$13.8_{4.8}$	$28.0_{8.2}$	$56.4_{12.1}$
	LongDocURL	$3.6_{2.1}$	$7.2_{4.5}$	$14.2_{8.1}$	$28.6_{14.4}$	$55.3_{18.3}$
	SlideVQA	$7.5_{0.9}$	$16.2_{2.1}$	$33.2_{2.7}$	$67.2_{3.5}$	$135.2_{5.2}$

Experimental Results

1. All models struggle, but closed-source models perform better.
   - Gemini-2.5-Pro is the best.

2. Models can generalize to longer context lengths.
   - Qwen2.5-VL-32B
   - Ovis2-34B



	VRAG					NIAH					ICL				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4o	80.5	74.7	71.8	74.2	67.3	79.6	73.8	67.5	65.4	57.1	99.0	98.2	96.0	92.4	88.4
Claude-3.7-Sonnet	84.9	81.8	66.7	67.6	68.8	63.1	61.2	54.1	N/A	N/A	97.0	94.2	N/A	N/A	N/A
Gemini-2.0-Flash	64.9	64.2	59.5	59.0	60.3	76.8	74.1	69.7	64.6	60.9	99.0	97.8	97.5	93.8	87.5
Gemini-2.0-Flash-T	67.0	68.5	66.7	67.0	64.4	80.8	79.2	76.2	68.7	64.8	99.5	97.8	96.2	92.5	88.2
Gemini-2.5-Flash	69.8	69.3	65.1	68.6	70.6	84.1	81.5	79.8	76.4	72.5	98.5	98.5	96.5	94.0	88.0
Gemini-2.5-Pro	79.8	80.9	79.9	80.8	82.7	84.7	82.7	79.8	76.0	73.4	99.5	98.5	97.2	95.0	94.2
Qwen2-VL-72B	64.3	64.0	60.1	56.1	46.9	63.9	61.6	57.4	51.5	38.9	98.5	94.5	91.0	80.8	80.8
Qwen2.5-VL-7B	50.1	48.7	43.2	36.8	31.6	57.3	53.0	47.7	39.5	33.2	95.6	91.5	78.5	57.2	46.2
Qwen2.5-VL-32B	67.8	69.1	65.5	61.9	64.6	61.9	61.1	58.5	53.7	41.6	97.5	91.7	77.0	51.2	41.2
Qwen2.5-VL-72B	67.6	67.7	64.0	54.3	50.3	68.3	63.5	61.9	55.8	43.1	98.5	95.5	92.8	74.2	73.0
InternVL2.5-26B	56.6	53.3	48.1	50.0	47.9	67.8	63.1	55.5	52.2	43.8	98.5	89.2	85.0	72.5	54.0
InternVL3-8B	52.3	51.3	45.8	40.3	36.3	62.6	57.8	51.8	49.7	42.4	97.6	87.2	75.0	61.8	8.5
InternVL3-14B	57.5	55.3	52.3	52.8	50.0	69.5	65.1	58.2	55.8	48.3	96.5	87.7	80.0	65.8	53.0
InternVL3-38B	65.7	60.8	52.2	50.4	40.3	70.5	66.4	62.5	57.0	52.0	99.5	95.0	88.5	77.5	65.2
Ovis2-8B	52.3	48.0	47.1	47.9	42.9	61.3	57.9	54.2	41.2	35.8	94.5	44.4	7.8	4.0	1.0
Ovis2-16B	56.2	51.2	49.7	49.2	41.3	67.3	62.7	56.5	48.7	40.7	96.6	91.2	73.2	66.0	36.5
Ovis2-34B	63.4	61.5	55.5	57.2	45.7	65.7	60.4	57.0	52.9	40.0	98.5	89.5	79.2	71.0	65.2
Gemma3-12B	58.6	52.1	46.9	43.5	41.7	60.7	55.9	51.4	47.5	41.7	99.0	96.5	93.2	82.2	59.0
Gemma3-27B	64.8	62.1	58.8	57.5	51.5	66.3	61.2	56.2	51.9	44.6	98.0	94.8	93.5	83.8	73.8
Idefics3-8B	33.3	31.8	30.3	35.2	33.2	49.2	45.2	43.1	39.6	37.5	25.6	12.3	4.5	0.8	2.0
Phi-4-Multimodal	36.3	37.3	35.4	32.9	25.5	48.8	44.6	41.1	36.7	34.9	82.3	42.5	12.0	2.8	2.2
NVILA-Lite-8B	43.2	41.6	41.8	35.8	16.3	52.7	47.8	43.6	36.8	29.0	93.1	73.6	47.0	20.5	2.8
Pixtral-12B	53.6	51.0	47.9	45.9	43.8	56.3	54.2	50.2	45.2	40.9	95.0	90.0	86.0	53.2	49.8

	Summ					DocVQA					Avg.				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4o	25.1	31.1	34.3	41.0	42.4	67.8	70.5	67.2	62.9	59.2	70.4	69.7	67.4	67.2	62.9
Claude-3.7-Sonnet	27.6	34.6	34.9	34.5	37.5	56.7	52.0	43.1	48.5	N/A	65.9	64.8	N/A	N/A	N/A
Gemini-2.0-Flash	24.4	27.1	30.1	30.6	35.9	58.7	55.4	59.4	53.8	53.6	64.8	63.7	63.2	60.3	59.6
Gemini-2.0-Flash-T	27.7	37.9	44.3	53.0	61.2	68.1	68.8	69.9	64.3	63.7	68.6	70.4	70.6	69.1	68.5
Gemini-2.5-Flash	29.2	39.4	45.9	55.3	62.4	67.5	66.9	68.6	62.5	59.3	69.8	71.1	71.2	71.4	70.5
Gemini-2.5-Pro	32.0	42.8	48.1	58.0	65.3	71.5	70.0	70.8	69.2	70.4	73.5	75.0	75.2	75.8	77.2
Qwen2-VL-72B	25.1	29.2	32.7	37.6	39.1	69.2	65.7	66.4	60.9	53.8	64.2	63.0	61.5	57.4	51.9
Qwen2.5-VL-7B	23.5	29.1	30.8	32.7	39.3	60.7	57.1	57.2	50.7	40.2	57.4	55.9	51.5	43.4	38.1
Qwen2.5-VL-32B	22.8	26.3	25.8	23.0	25.2	67.8	66.0	65.8	58.4	53.6	63.6	62.9	58.5	49.7	45.2
Qwen2.5-VL-72B	20.5	26.9	31.1	38.0	28.5	71.4	67.5	65.8	57.3	48.7	65.2	64.2	63.1	55.9	48.7
InternVL2.5-26B	19.1	23.8	26.3	27.8	29.5	53.5	47.6	51.4	44.6	32.8	59.1	55.4	53.3	49.4	41.6
InternVL3-8B	22.2	28.6	32.5	36.6	40.8	58.1	53.7	55.3	48.7	42.6	58.5	55.7	52.1	47.4	34.1
InternVL3-14B	22.3	25.6	27.2	30.3	35.8	63.3	54.1	57.5	50.0	39.4	61.8	57.5	55.1	50.9	45.3
InternVL3-38B	20.7	24.8	33.1	38.4	43.6	66.3	63.8	62.9	52.2	47.9	64.5	62.1	59.9	55.1	49.8
Ovis2-8B	23.0	29.3	30.5	32.9	28.3	59.1	49.3	42.3	30.3	10.9	58.0	45.8	36.4	31.3	23.8
Ovis2-16B	25.3	30.0	33.5	37.0	39.3	66.5	61.2	48.5	35.4	19.3	62.4	59.3	52.3	47.3	35.4
Ovis2-34B	23.5	29.8	35.7	39.6	41.6	59.9	55.2	45.2	33.6	23.5	62.2	59.3	54.5	50.9	43.2
Gemma3-12B	21.0	24.0	25.2	26.1	28.0	42.7	43.2	43.2	39.2	41.3	56.4	54.4	52.0	47.7	42.3
Gemma3-27B	22.9	28.5	32.0	35.5	40.7	49.7	49.7	45.5	46.2	45.6	60.4	59.3	57.2	55.0	51.2
Idefics3-8B	15.7	20.4	19.2	21.8	17.7	46.3	37.1	42.0	26.4	17.3	34.0	29.4	27.8	24.7	21.5
Phi-4-Multimodal	12.3	17.4	17.5	18.8	15.9	44.5	45.5	47.9	41.7	26.0	44.8	37.5	30.8	26.6	20.9
NVILA-Lite-8B	12.8	15.3	19.3	19.9	23.3	30.8	32.4	25.8	21.6	20.6	46.5	42.1	35.5	26.9	18.4
Pixtral-12B	22.7	29.6	33.5	36.7	38.5	55.0	48.1	44.4	38.7	32.4	56.5	54.6	52.4	43.9	41.1

Experimental Results

3. Reasoning can improve multimodal long-context ability.

- Gemini-2.0-Flash-T
- Gemini2.5-Flash&Pro

4. Different models exhibit different strengths.

- Qwen2.5-VL-32B better on VRAG
- InternVL3-38B better on NIAH

<u>Need Comprehensive Evaluation!</u>

	VRAG					NIAH					ICL				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4o	80.5	74.7	71.8	74.2	67.3	79.6	73.8	67.5	65.4	57.1	99.0	98.2	96.0	92.4	88.4
Claude-3.7-Sonnet	84.9	81.8	66.7	67.6	68.8	63.1	61.2	54.1	N/A	N/A	97.0	94.2	N/A	N/A	N/A
Gemini-2.0-Flash	64.9	64.2	59.5	59.0	60.3	76.8	74.1	69.7	64.6	60.9	99.0	97.8	97.5	93.8	87.5
Gemini-2.0-Flash-T	67.0	68.5	66.7	67.0	64.4	80.8	79.2	76.2	68.7	64.8	99.5	97.8	96.2	92.5	88.2
Gemini-2.5-Flash	69.8	69.3	65.1	68.6	70.6	84.1	81.5	79.8	76.4	72.5	98.5	98.5	96.5	94.0	88.0
Gemini-2.5-Pro	79.8	80.9	79.9	80.8	82.7	84.7	82.7	79.8	76.0	73.4	99.5	98.5	97.2	95.0	94.2
Qwen2-VL-72B	64.3	64.0	60.1	56.1	46.9	63.9	61.6	57.4	51.5	38.9	98.5	94.5	91.0	80.8	80.8
Qwen2.5-VL-7B	50.1	48.7	43.2	36.8	31.6	57.3	53.0	47.7	39.5	33.2	95.6	91.5	78.5	57.2	46.2
Qwen2.5-VL-32B	67.8	69.1	65.5	61.9	64.6	61.9	61.1	58.5	53.7	41.6	97.5	91.7	77.0	51.2	41.2
Qwen2.5-VL-72B	67.6	67.7	64.0	54.3	50.3	68.3	63.5	61.9	55.8	43.1	98.5	95.5	92.8	74.2	73.0
InternVL2.5-26B	56.6	53.3	48.1	50.0	47.9	67.8	63.1	55.5	52.2	43.8	98.5	89.2	85.0	72.5	54.0
InternVL3-8B	52.3	51.3	45.8	40.3	36.3	62.6	57.8	51.8	49.7	42.4	97.6	87.2	75.0	61.8	8.5
InternVL3-14B	57.5	55.3	52.3	52.8	50.0	69.5	65.1	58.2	55.8	48.3	96.5	87.7	80.0	65.8	53.0
InternVL3-38B	65.7	60.8	52.2	50.4	40.3	70.5	66.4	62.5	57.0	52.0	99.5	95.0	88.5	77.5	65.2
Ovis2-8B	52.3	48.0	47.1	47.9	42.9	61.3	57.9	54.2	41.2	35.8	94.5	44.4	7.8	4.0	1.0
Ovis2-16B	56.2	51.2	49.7	49.2	41.3	67.3	62.7	56.5	48.7	40.7	96.6	91.2	73.2	66.0	36.5
Ovis2-34B	63.4	61.5	55.5	57.2	45.7	65.7	60.4	57.0	52.9	40.0	98.5	89.5	79.2	71.0	65.2
Gemma3-12B	58.6	52.1	46.9	43.5	41.7	60.7	55.9	51.4	47.5	41.7	99.0	96.5	93.2	82.2	59.0
Gemma3-27B	64.8	62.1	58.8	57.5	51.5	66.3	61.2	56.2	51.9	44.6	98.0	94.8	93.5	83.8	73.8
Idefics3-8B	33.3	31.8	30.3	35.2	33.2	49.2	45.2	43.1	39.6	37.5	25.6	12.3	4.5	0.8	2.0
Phi-4-Multimodal	36.3	37.3	35.4	32.9	25.5	48.8	44.6	41.1	36.7	34.9	82.3	42.5	12.0	2.8	2.2
NVILA-Lite-8B	43.2	41.6	41.8	35.8	16.3	52.7	47.8	43.6	36.8	29.0	93.1	73.6	47.0	20.5	2.8
Pixtral-12B	53.6	51.0	47.9	45.9	43.8	56.3	54.2	50.2	45.2	40.9	95.0	90.0	86.0	53.2	49.8

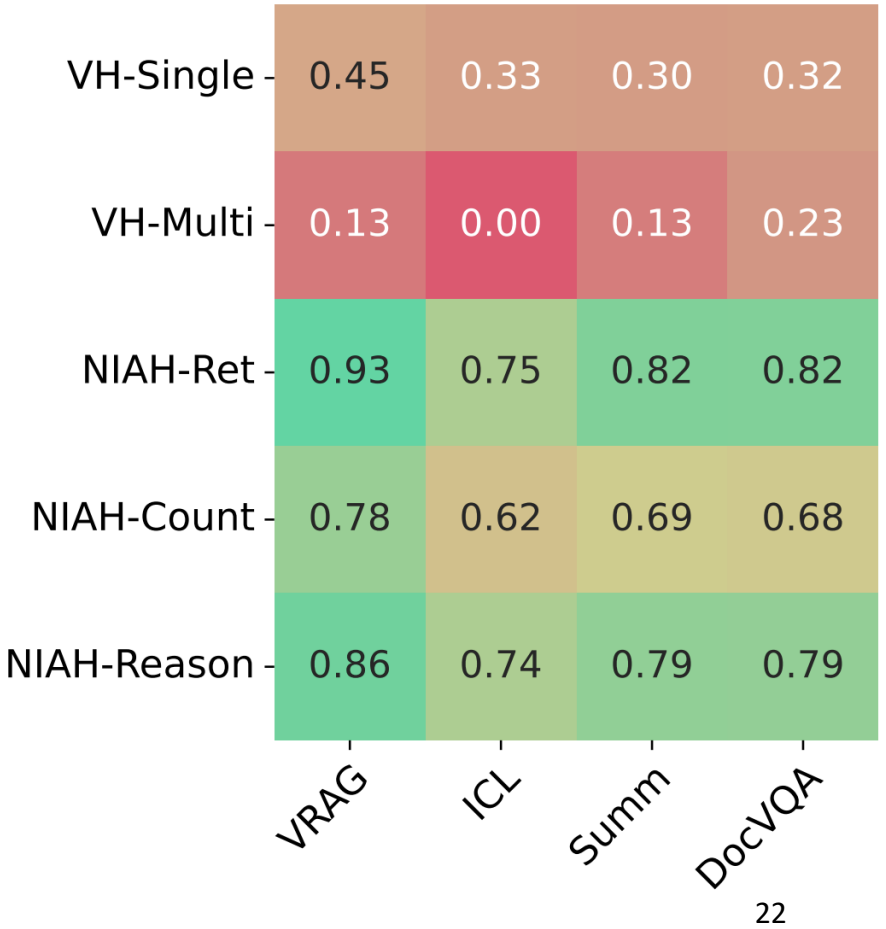
	Summ					DocVQA					Avg.				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4o	25.1	31.1	34.3	41.0	42.4	67.8	70.5	67.2	62.9	59.2	70.4	69.7	67.4	67.2	62.9
Claude-3.7-Sonnet	27.6	34.6	34.9	34.5	37.5	56.7	52.0	43.1	48.5	N/A	65.9	64.8	N/A	N/A	N/A
Gemini-2.0-Flash	24.4	27.1	30.1	30.6	35.9	58.7	55.4	59.4	53.8	53.6	64.8	63.7	63.2	60.3	59.6
Gemini-2.0-Flash-T	27.7	37.9	44.3	53.0	61.2	68.1	68.8	69.9	64.3	63.7	68.6	70.4	70.6	69.1	68.5
Gemini-2.5-Flash	29.2	39.4	45.9	55.3	62.4	67.5	66.9	68.6	62.5	59.3	69.8	71.1	71.2	71.4	70.5
Gemini-2.5-Pro	32.0	42.8	48.1	58.0	65.3	71.5	70.0	70.8	69.2	70.4	73.5	75.0	75.2	75.8	77.2
Qwen2-VL-72B	25.1	29.2	32.7	37.6	39.1	69.2	65.7	66.4	60.9	53.8	64.2	63.0	61.5	57.4	51.9
Qwen2.5-VL-7B	23.5	29.1	30.8	32.7	39.3	60.7	57.1	57.2	50.7	40.2	57.4	55.9	51.5	43.4	38.1
Qwen2.5-VL-32B	22.8	26.3	25.8	23.0	25.2	67.8	66.0	65.8	58.4	53.6	63.6	62.9	58.5	49.7	45.2
Qwen2.5-VL-72B	20.5	26.9	31.1	38.0	28.5	71.4	67.5	65.8	57.3	48.7	65.2	64.2	63.1	55.9	48.7
InternVL2.5-26B	19.1	23.8	26.3	27.8	29.5	53.5	47.6	51.4	44.6	32.8	59.1	55.4	53.3	49.4	41.6
InternVL3-8B	22.2	28.6	32.5	36.6	40.8	58.1	53.7	55.3	48.7	42.6	58.5	55.7	52.1	47.4	34.1
InternVL3-14B	22.3	25.6	27.2	30.3	35.8	63.3	54.1	57.5	50.0	39.4	61.8	57.5	55.1	50.9	45.3
InternVL3-38B	20.7	24.8	33.1	38.4	43.6	66.3	63.8	62.9	52.2	47.9	64.5	62.1	59.9	55.1	49.8
Ovis2-8B	23.0	29.3	30.5	32.9	28.3	59.1	49.3	42.3	30.3	10.9	58.0	45.8	36.4	31.3	23.8
Ovis2-16B	25.3	30.0	33.5	37.0	39.3	66.5	61.2	48.5	35.4	19.3	62.4	59.3	52.3	47.3	35.4
Ovis2-34B	23.5	29.8	35.7	39.6	41.6	59.9	55.2	45.2	33.6	23.5	62.2	59.3	54.5	50.9	43.2
Gemma3-12B	21.0	24.0	25.2	26.1	28.0	42.7	43.2	43.2	39.2	41.3	56.4	54.4	52.0	47.7	42.3
Gemma3-27B	22.9	28.5	32.0	35.5	40.7	49.7	49.7	45.5	46.2	45.6	60.4	59.3	57.2	55.0	51.2
Idefics3-8B	15.7	20.4	19.2	21.8	17.7	46.3	37.1	42.0	26.4	17.3	34.0	29.4	27.8	24.7	21.5
Phi-4-Multimodal	12.3	17.4	17.5	18.8	15.9	44.5	45.5	47.9	41.7	26.0	44.8	37.5	30.8	26.6	20.9
NVILA-Lite-8B	12.8	15.3	19.3	19.9	23.3	30.8	32.4	25.8	21.6	20.6	46.5	42.1	35.5	26.9	18.4
Pixtral-12B	22.7	29.6	33.5	36.7	38.5	55.0	48.1	44.4	38.7	32.4	56.5	54.6	52.4	43.9	41.1

Evaluation Suggestions

Can NIAH Tasks Reflect LCVLM's Overall Long-Context Ability?

No!

- Visual Haystack: Low Correlation
- MM-NIAH: < 0.8



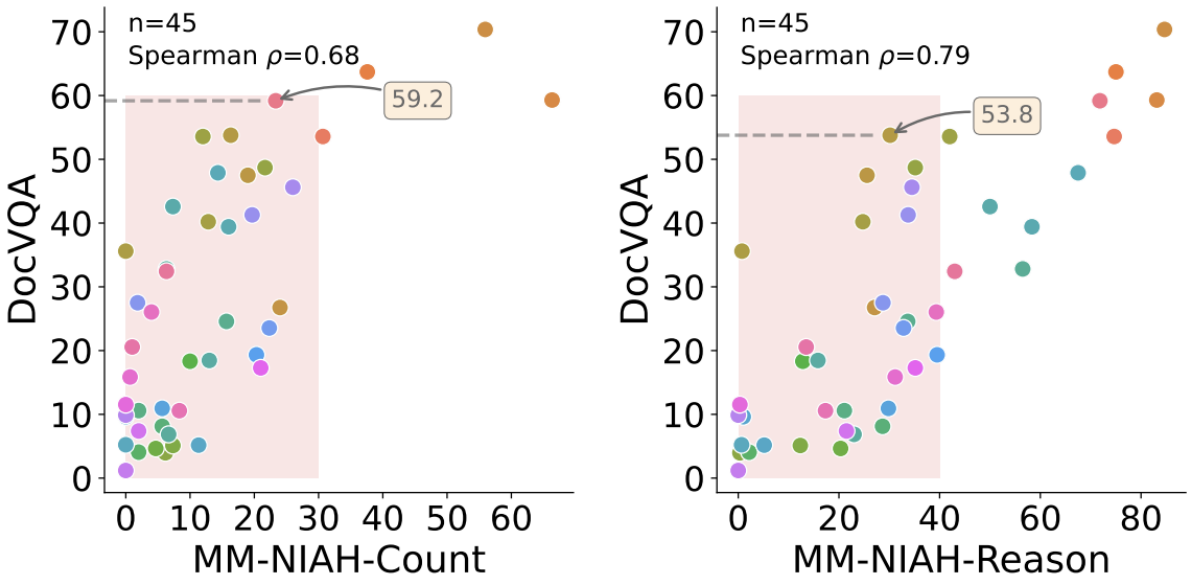
Evaluation Suggestions

NIAH tasks are hard for current LVLMs!

VH-Multi

	8k	16k	32k	64k	128k
GPT-4o	70.3	65.0	63.3	61.0	50.7
Gemini-2.5-Pro	76.5	72.9	68.9	70.1	67.7
Qwen2.5-VL-7B	54.8	53.7	54.0	55.2	54.8
Qwen2.5-VL-32B	57.5	56.0	55.8	56.2	53.7
Qwen2.5-VL-72B	64.3	57.7	54.5	54.0	55.0
Gemma3-4B	52.7	59.0	52.8	56.5	52.7
Gemma3-12B	56.3	51.3	52.5	51.0	53.2
Gemma3-27B	57.2	53.5	56.3	60.3	57.0

Random guess yields 50% accuracy, highlighting its difficulty.



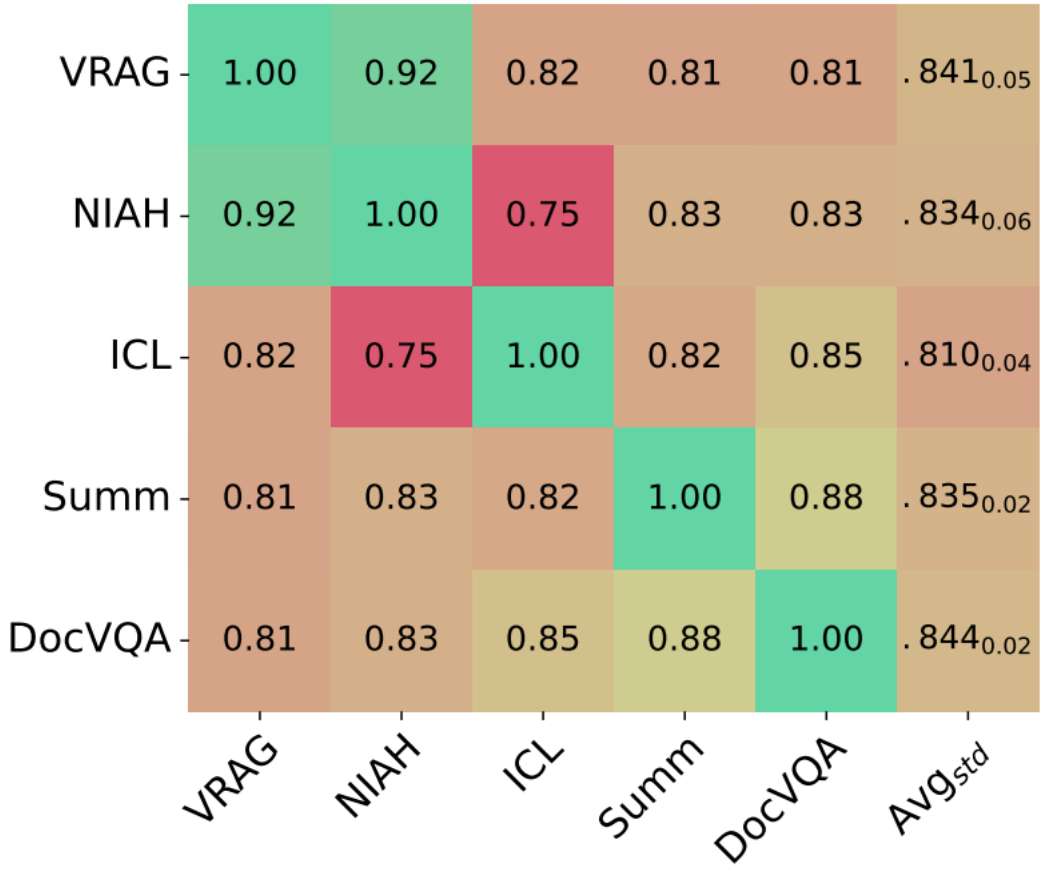
Scores below 30 and 40, poor separability between models

Evaluation Suggestions

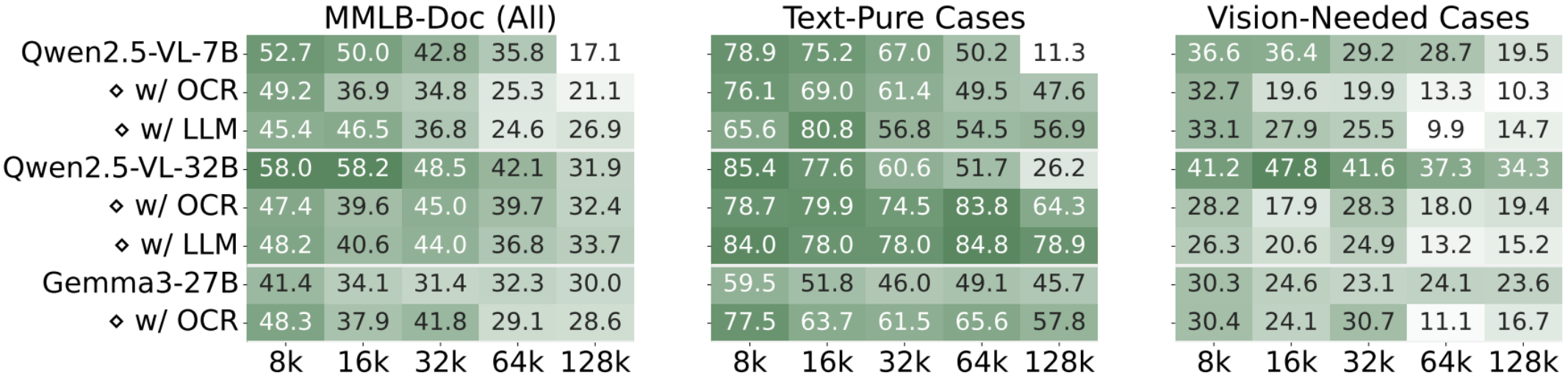
Correlations Across Categories are not strong, consistently < 0.85

We need comprehensive evaluation!

Long-document VQA is a reliable proxy for fast iteration.



Error Analysis & Case Studies



MMLongBench-Doc

- Use OCR-extracted text instead of PDF images (w/ OCR)
- Use instruction version of Qwen2.5-7B and Qwen2.5-32B (w/ LLM; text-only models)
- Categorize examples according to the answer sources: text-pure and vision-needed.

Error Analysis & Case Studies

MMLB-Doc (All)

	8k	16k	32k	64k	128k
Qwen2.5-VL-7B	52.7	50.0	42.8	35.8	17.1
◇ w/ OCR	49.2	36.9	34.8	25.3	21.1
◇ w/ LLM	45.4	46.5	36.8	24.6	26.9
Qwen2.5-VL-32B	58.0	58.2	48.5	42.1	31.9
◇ w/ OCR	47.4	39.6	45.0	39.7	32.4
◇ w/ LLM	48.2	40.6	44.0	36.8	33.7
Gemma3-27B	41.4	34.1	31.4	32.3	30.0
◇ w/ OCR	48.3	37.9	41.8	29.1	28.6

Text-Pure Cases

	8k	16k	32k	64k	128k
Qwen2.5-VL-7B	78.9	75.2	67.0	50.2	11.3
◇ w/ OCR	76.1	69.0	61.4	49.5	47.6
◇ w/ LLM	65.6	80.8	56.8	54.5	56.9
Qwen2.5-VL-32B	85.4	77.6	60.6	51.7	26.2
◇ w/ OCR	78.7	79.9	74.5	83.8	64.3
◇ w/ LLM	84.0	78.0	78.0	84.8	78.9
Gemma3-27B	59.5	51.8	46.0	49.1	45.7
◇ w/ OCR	77.5	63.7	61.5	65.6	57.8

Vision-Needed Cases

	8k	16k	32k	64k	128k
Qwen2.5-VL-7B	36.6	36.4	29.2	28.7	19.5
◇ w/ OCR	32.7	19.6	19.9	13.3	10.3
◇ w/ LLM	33.1	27.9	25.5	9.9	14.7
Qwen2.5-VL-32B	41.2	47.8	41.6	37.3	34.3
◇ w/ OCR	28.2	17.9	28.3	18.0	19.4
◇ w/ LLM	26.3	20.6	24.9	13.2	15.2
Gemma3-27B	30.3	24.6	23.1	24.1	23.6
◇ w/ OCR	30.4	24.1	30.7	11.1	16.7

Findings:

- No clear winner between OCR-extracted text and PDF images
- PDF images lead to higher scores in vision-needed cases
- OCR yields better performance in text-pure cases (OCR tool > VLM OCR)

VLM's OCR ability in long context is a bottleneck!

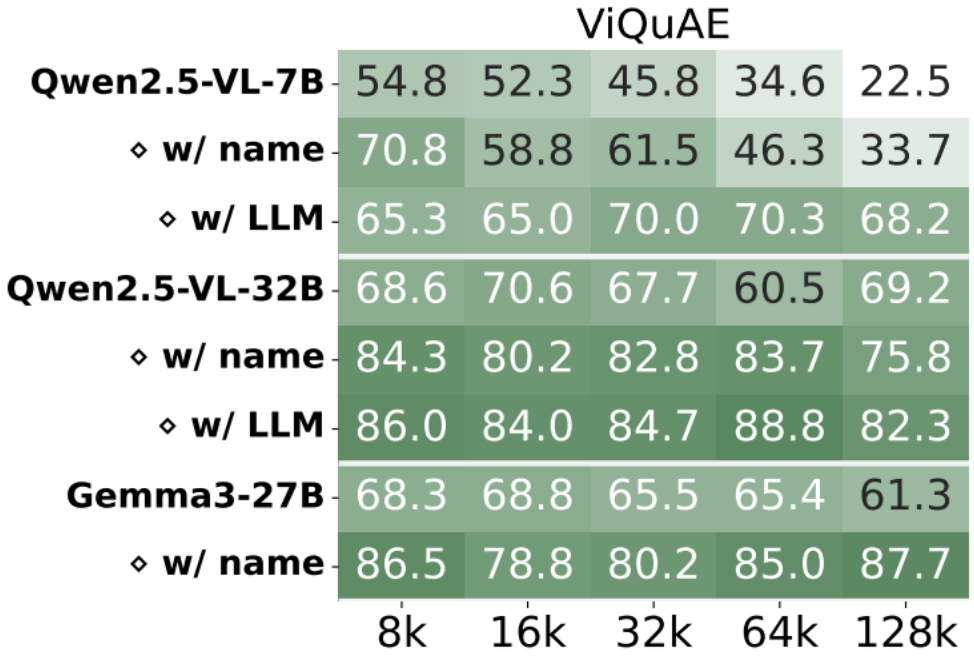
Error Analysis & Case Studies

Visual RAG (InfoSeek)

- Replace the image with entity name
- Use instruction version of Qwen2.5-7B and Qwen2.5-32B (w/ LLM; text-only models)

Findings:

- All models improves when we directly provide entity name
- Text-only models perform better



Cross-modal retrieval in long context is a bottleneck! (from entity image to its attribute)

More Experiments & Analysis (YaRN)

VRAG	8k	16k	32k	64k	128k
Qwen2.5-VL-3B	43.9	38.6	35.8	32.7	9.8
⋄ w/ Yarn	43.7	39.3	33.5	34.6	30.0
Qwen2.5-VL-7B	50.1	48.7	43.2	36.8	31.6
⋄ w/ Yarn	45.6	42.6	42.2	38.9	32.5
Qwen2.5-VL-32B	67.8	69.1	65.5	61.9	64.6
⋄ w/ Yarn	68.2	66.8	64.4	63.1	54.2
Qwen2.5-VL-72B	67.6	67.7	64.0	54.3	50.3
⋄ w/ Yarn	68.6	66.7	63.7	61.9	53.7

NIAH	8k	16k	32k	64k	128k
Qwen2.5-VL-3B	54.1	50.8	45.0	38.7	21.6
⋄ w/ Yarn	56.7	52.8	50.3	48.3	39.7
Qwen2.5-VL-7B	57.3	53.0	47.7	39.5	33.2
⋄ w/ Yarn	59.4	57.2	55.0	50.9	44.6
Qwen2.5-VL-32B	61.9	61.1	58.5	53.7	41.6
⋄ w/ Yarn	64.2	61.1	58.6	56.6	51.1
Qwen2.5-VL-72B	68.3	63.5	61.9	55.8	43.1
⋄ w/ Yarn	71.1	66.1	63.2	59.5	55.9

ICL	8k	16k	32k	64k	128k
Qwen2.5-VL-3B	95.0	69.9	19.5	7.5	9.0
⋄ w/ Yarn	80.4	47.0	12.2	3.5	5.0
Qwen2.5-VL-7B	95.6	91.5	78.5	57.2	46.2
⋄ w/ Yarn	98.5	91.7	80.5	63.8	51.5
Qwen2.5-VL-32B	97.5	91.7	77.0	51.2	41.2
⋄ w/ Yarn	97.5	82.7	59.0	43.0	37.5
Qwen2.5-VL-72B	98.5	95.5	92.8	74.2	73.0
⋄ w/ Yarn	99.0	96.2	93.8	79.8	68.0

Summ	8k	16k	32k	64k	128k
Qwen2.5-VL-3B	18.8	23.2	24.9	27.1	30.2
⋄ w/ Yarn	18.5	21.4	24.5	28.6	31.5
Qwen2.5-VL-7B	23.5	29.1	30.8	32.7	39.3
⋄ w/ Yarn	21.5	26.7	27.8	32.1	37.1
Qwen2.5-VL-32B	22.8	26.3	25.8	23.0	25.2
⋄ w/ Yarn	21.3	23.4	24.4	23.5	23.5
Qwen2.5-VL-72B	20.5	26.9	31.1	38.0	28.5
⋄ w/ Yarn	20.1	24.1	23.9	26.5	25.6

DocVQA	8k	16k	32k	64k	128k
Qwen2.5-VL-3B	55.5	52.0	51.7	45.0	35.6
⋄ w/ Yarn	53.5	48.3	54.9	45.2	44.5
Qwen2.5-VL-7B	60.7	57.1	57.2	50.7	40.2
⋄ w/ Yarn	60.2	56.3	59.0	55.1	49.3
Qwen2.5-VL-32B	67.8	66.0	65.8	58.4	53.6
⋄ w/ Yarn	63.4	62.4	65.1	60.0	55.3
Qwen2.5-VL-72B	71.4	67.5	65.8	57.3	48.7
⋄ w/ Yarn	71.0	67.2	63.3	57.9	49.1

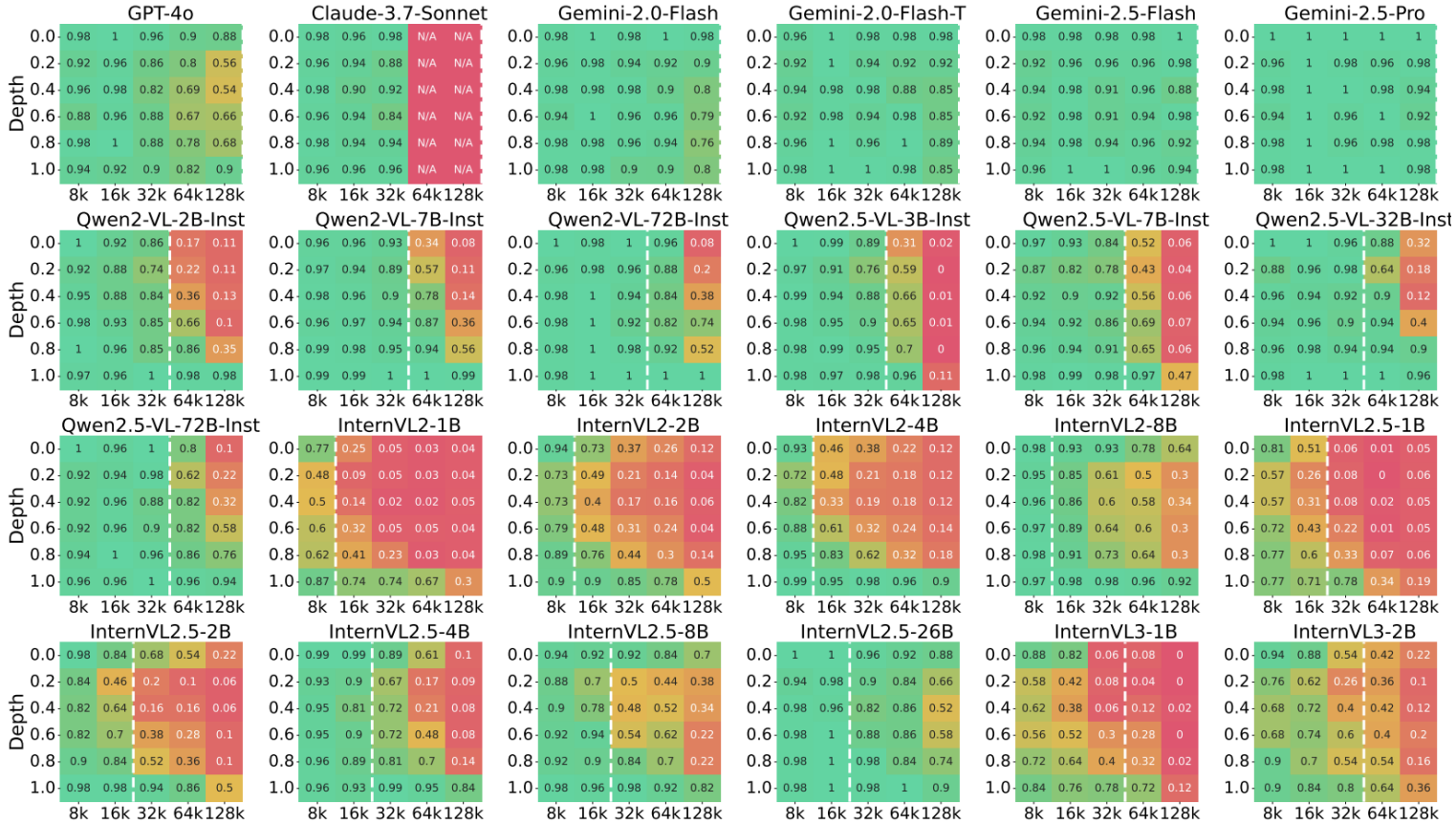
Avg.	8k	16k	32k	64k	128k
Qwen2.5-VL-3B	53.5	46.9	35.4	30.2	21.2
⋄ w/ Yarn	50.5	41.8	35.1	32.0	30.2
Qwen2.5-VL-7B	57.4	55.9	51.5	43.4	38.1
⋄ w/ Yarn	57.0	54.9	52.9	48.2	43.0
Qwen2.5-VL-32B	63.6	62.9	58.5	49.7	45.2
⋄ w/ Yarn	62.9	59.3	54.3	49.3	44.3
Qwen2.5-VL-72B	65.2	64.2	63.1	55.9	48.7
⋄ w/ Yarn	65.9	64.1	61.6	57.1	50.5

- Performance on shorter input length hurts
- Only 3B model improves a lot at 128K
- 32B and 72B only fluctuate

More Experiments & Analysis

- Full model evaluation (46 models)
- V2PE on InternVL2
- Lost in the middle
- …

Lost in the middle (partial)



Conclusions & Future Work

- First comprehensive benchmark for long-context vision-language models (LCVLMs)
- Rigorous, extensible foundation for diagnosing the strengths and weaknesses of frontier LCVLMs
- Evaluation on 46 models reveals with rich insights

Looking forward, we hope MMLONGBENCH will serve as a standard yardstick for the community to benchmark new LCVLMs

What's new

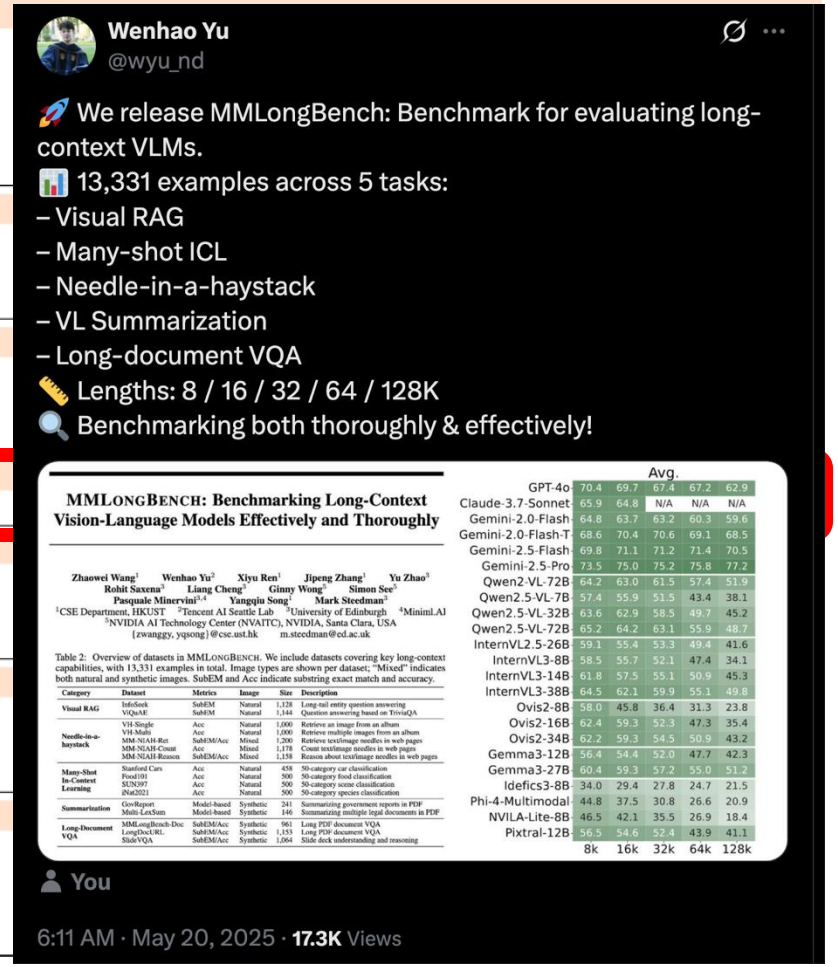
A few Vision-Language teams has MMLongBench:

- MiMo-VL 2508[1] (Xiaomi)
- Kimi-VL (Moonshot)
- Seed-VL (ByteDance)

Invited Talks:

- MLNLP[2]
- ResearchTrend[3]

Benchmark	Keye-VL-8B (Thinking)	GLM-4.1V-9B (Thinking)	MiMo-VL-7B-RL (Thinking)	MiMo-VL-7B-RL-2508 (Non-Thinking)	MiMo-VL-7B-RL-2508 (Thinking)
General					
$MMMU_{val}$					
V*					
$ZeroBench_{sub}$					
MM-IFEval					
$RefCOCO^{avg}_{val}$					
Doc & OCR					
ChartQA					
OCRBench					
$InfoVQA_{val}$					
Video					
Video-MME (w/o sub.)					
Video-MMMU					
VSI-Bench					
Long Context					
$MMLongBench_{avg}$ (32k)					
GUI					
ScreenSpot-v2					
CAGUI					
$AndroidControl_{avg}$					
Reasoning					
OlympiadBench					
LogicVista					
PhysReason					
Text					
SuperGPQA					
DROP					
AIME24					
AIME25					



[1] https://huggingface.co/XiaomiMiMo/MiMo-VL-7B-RL-2508
[2] https://www.bilibili.com/video/BV12Hu7zeE11/?share_source=copy_web&vd_source=e92a97b90c150ef916306fe0fdb9ee6f
[3] https://researchtrend.ai/social-events/researchtrend-connect-vlm-llmag-aifin

2. SubeventWriter: Iterative Sub-event Sequence Generation with Coherence Controller

Zhaowei Wang, Hongming Zhang, Tianqing Fang, Yangqiu Song, Ginny Y. Wong & Simon See

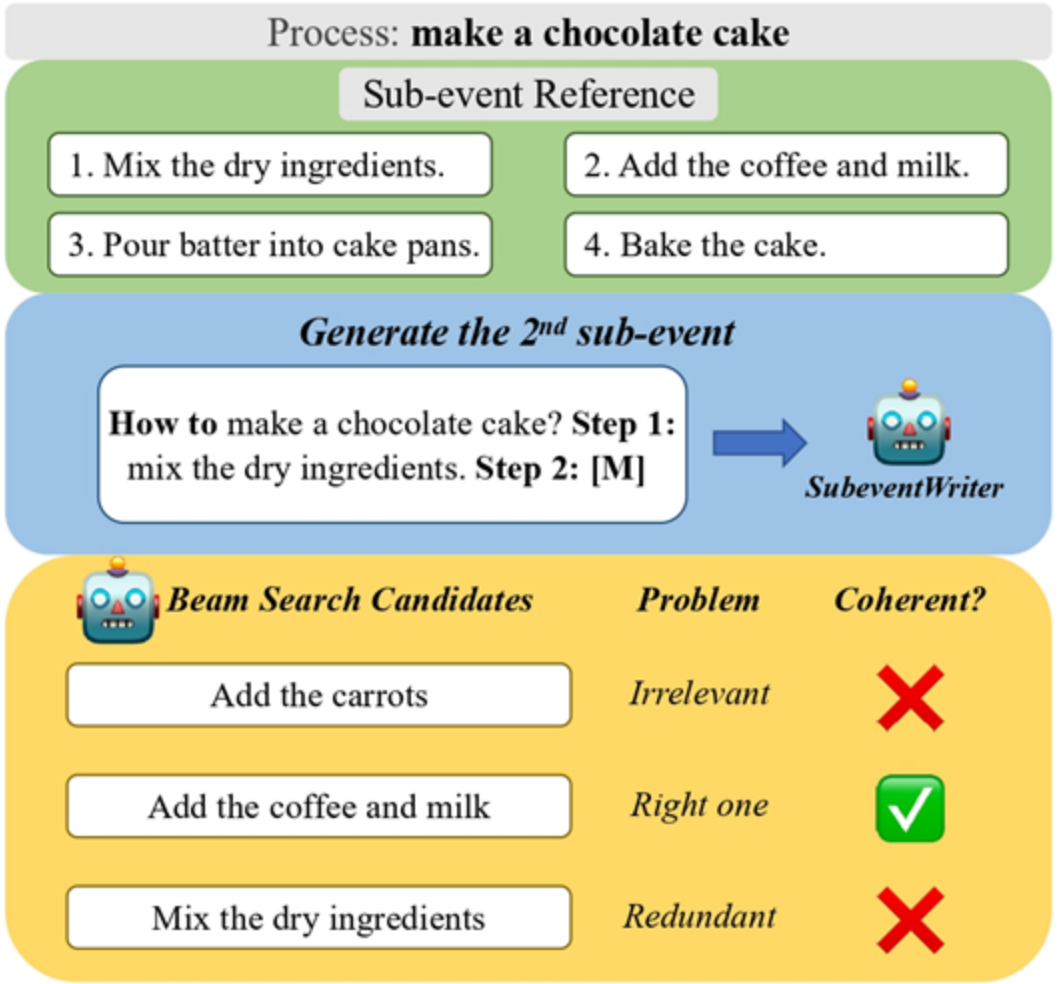
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

NVIDIA

Task Formulation

Given a process, can language models <u>plan</u> right steps?

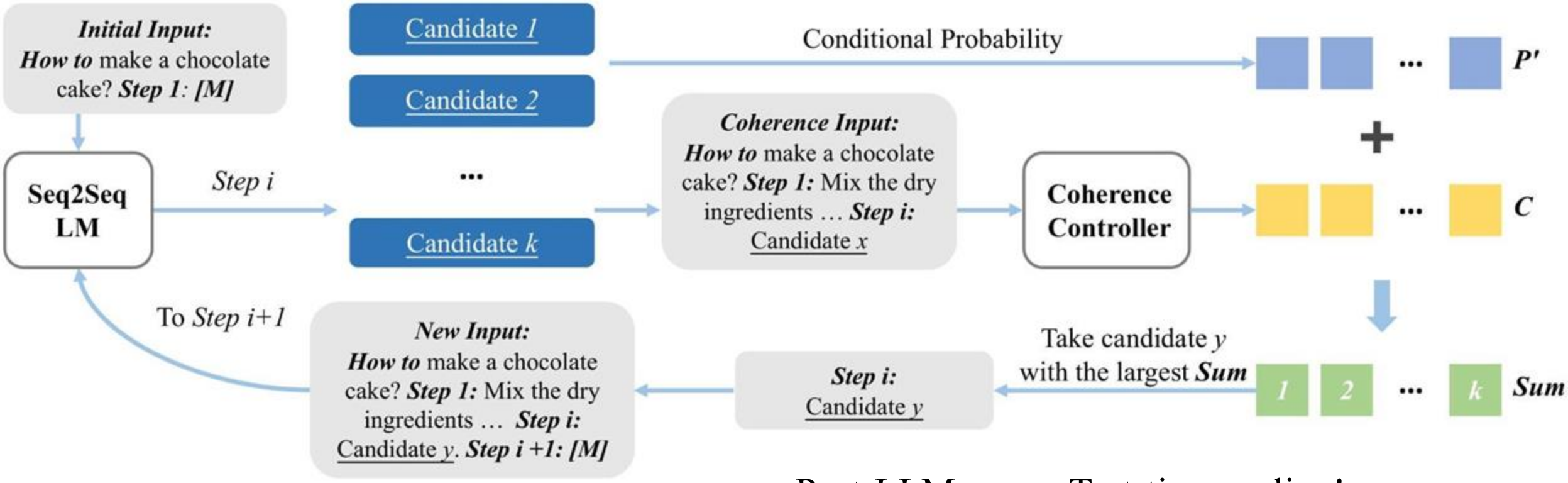
Our motivation is about coherence among steps when doing the planning.



Method

1. Iterative Event-level Decoding: a Seq2Seq Language Model generate steps one by one.

2. Coherence Controller: a masked language model generates coherence scores.



Post-LLM name: Test-time scaling!

Result

~83k processes from WikiHow[1]

Seq2Seq Models:
BART-base/large
T5-base/large/3b

Models	B-1	B-2	R-L	BERT	Δ_{B-1}	Δ_{B-2}
Zero-shot Large LM (GPT-J 6b)	13.88	0.33	16.47	45.43	-	-
Zero-shot Large LM (T5-11b)	20.14	0.76	14.11	54.55	-	-
Top-1 Similar Sequence (Glove)	16.31	0.99	11.63	57.24	-	-
Top-1 Similar Sequence (SBERT)	18.39	2.21	13.46	59.94	-	-
All-at-once Seq2Seq (BART-base)	21.01	4.52	18.83	58.79	-	-
All-at-once Seq2Seq (BART-large)	21.84	4.73	18.94	59.45	-	-
All-at-once Seq2Seq (T5-base)	20.33	5.63	20.22	52.15	-	-
All-at-once Seq2Seq (T5-large)	24.27	7.11	21.76	57.58	-	-
All-at-once Seq2Seq (T5-3b)	27.99	8.72	23.36	62.03	-	-
SubeventWriter (BART-base)	29.62	8.35	21.59	60.42	↑ 8.61	↑ 3.83
SubeventWriter (BART-large)	31.31	9.41	22.52	61.83	↑ 9.47	↑ **4.68**
SubeventWriter (T5-base)	30.74	8.89	22.44	61.81	↑ **10.41**	↑ 3.26
SubeventWriter (T5-large)	33.01	10.39	23.07	64.19	↑ 8.74	↑ 3.28
SubeventWriter (T5-3b)	**34.75**	**11.30**	**24.17**	**65.67**	↑ 6.76	↑ 2.58

[1] https://www.wikihow.com

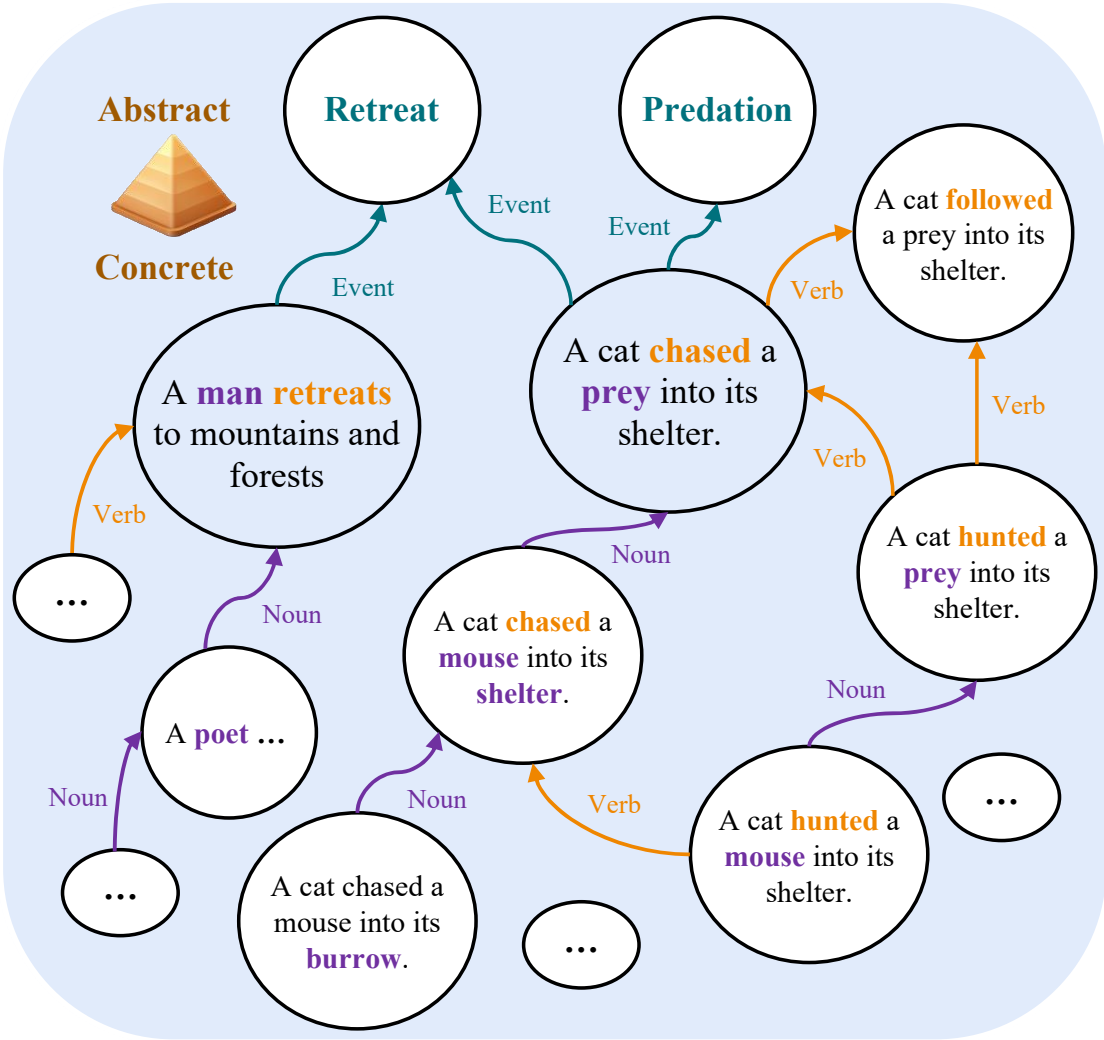
3. AbsInstruct: Eliciting Abstraction Ability from LLMs through Explanation Tuning with Plausibility Estimation

Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Y. Wong, & Simon See

Task Formulation

Entailment relation (Abstraction) between events

- Based on our previous work: AbsPyramid[1].
- We can know Event B given Event A, contextualized is-a relation
- Cognitive Study: K-Line Theory[2]

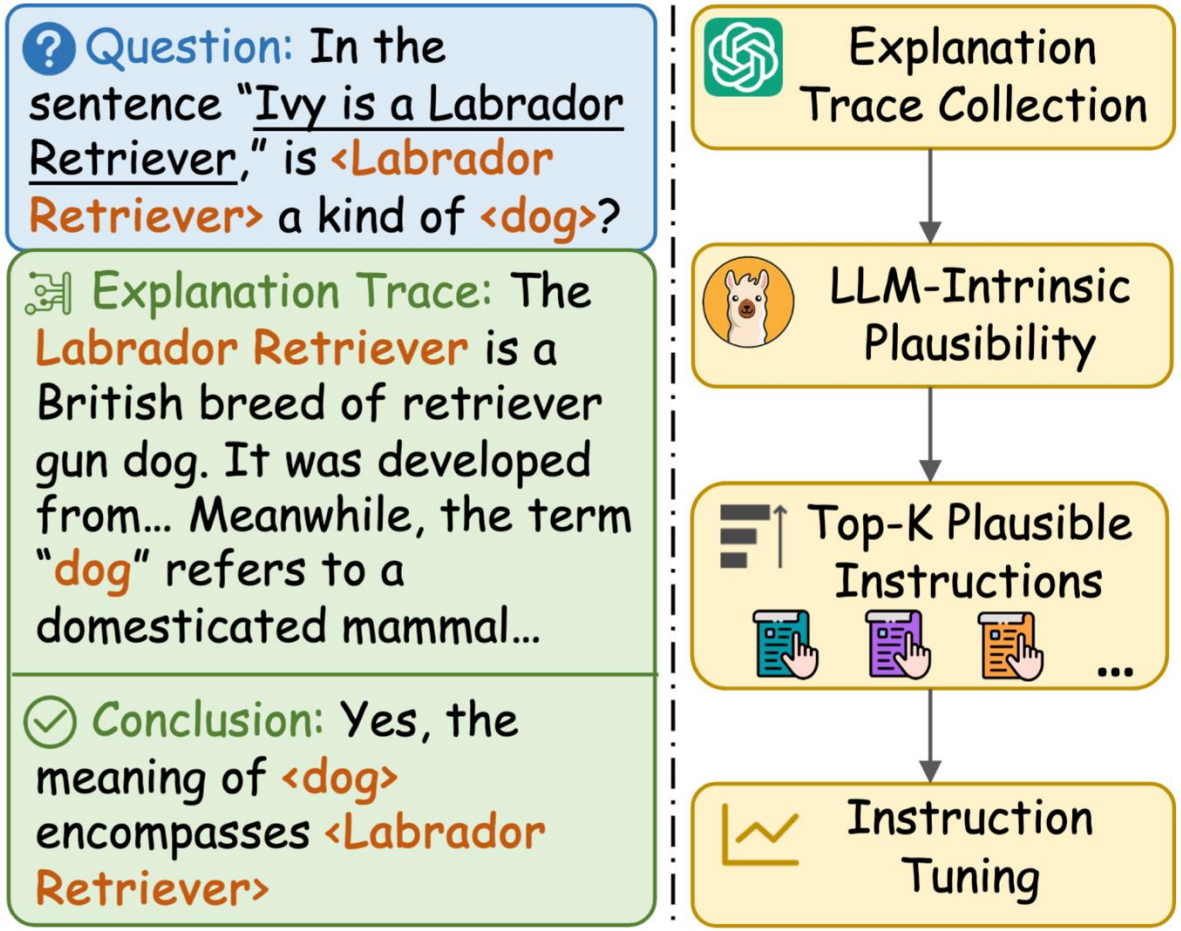


[1] AbsPyramid: Benchmarking the Abstraction Ability of Language Models with a Unified Entailment Graph
[2] Minsky, Marvin. "K-Lines: A theory of Memory." Cognitive science 4.2 (1980): 117-133.

Method

Enhance the abstraction ability of LLM with Instruction Tuning

- **General Domain Data:** Alpaca
- **Building Specific Domain Data**:
  - Data from AbsPyramid (221K)
  - Explanation Trace (effective)
  - Perplexity Filtering (efficient)



[2] Taori, Rohan, et al. "Stanford alpaca: An instruction-following llama model." 30 Jun. 2023

Method

❑ **Response Collection with Explanation**

1) **Explanation Step:** GPT4 generates the meaning of given words

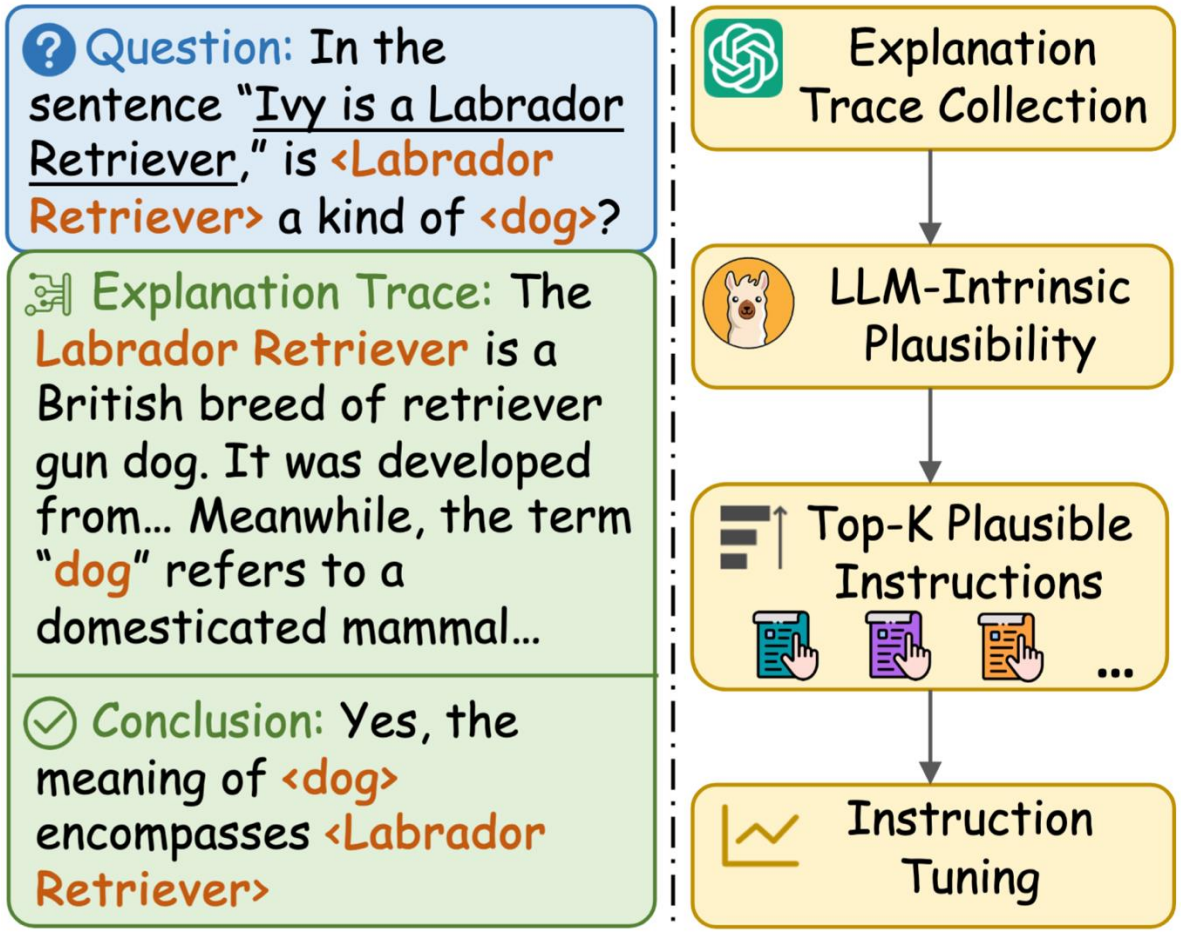
2) **Conclusion Step:** Yes, the meaning of [cpt] encompasses [ins]. / No, the meaning of [cpt] doesn't encompass [ins].

❑ **Perplexity Filtering**

1) LLMs gain knowledge during pre-training[3]

2) Only need to better elicit the abstraction knowledge?

3) Compute the perplexity of each abstraction example (This can show pre-train knowledge)



[3] Zhou, Chunting, et al. "Lima: Less is more for alignment." Advances in Neural Information Processing Systems 36 (2023): 55006-55021.

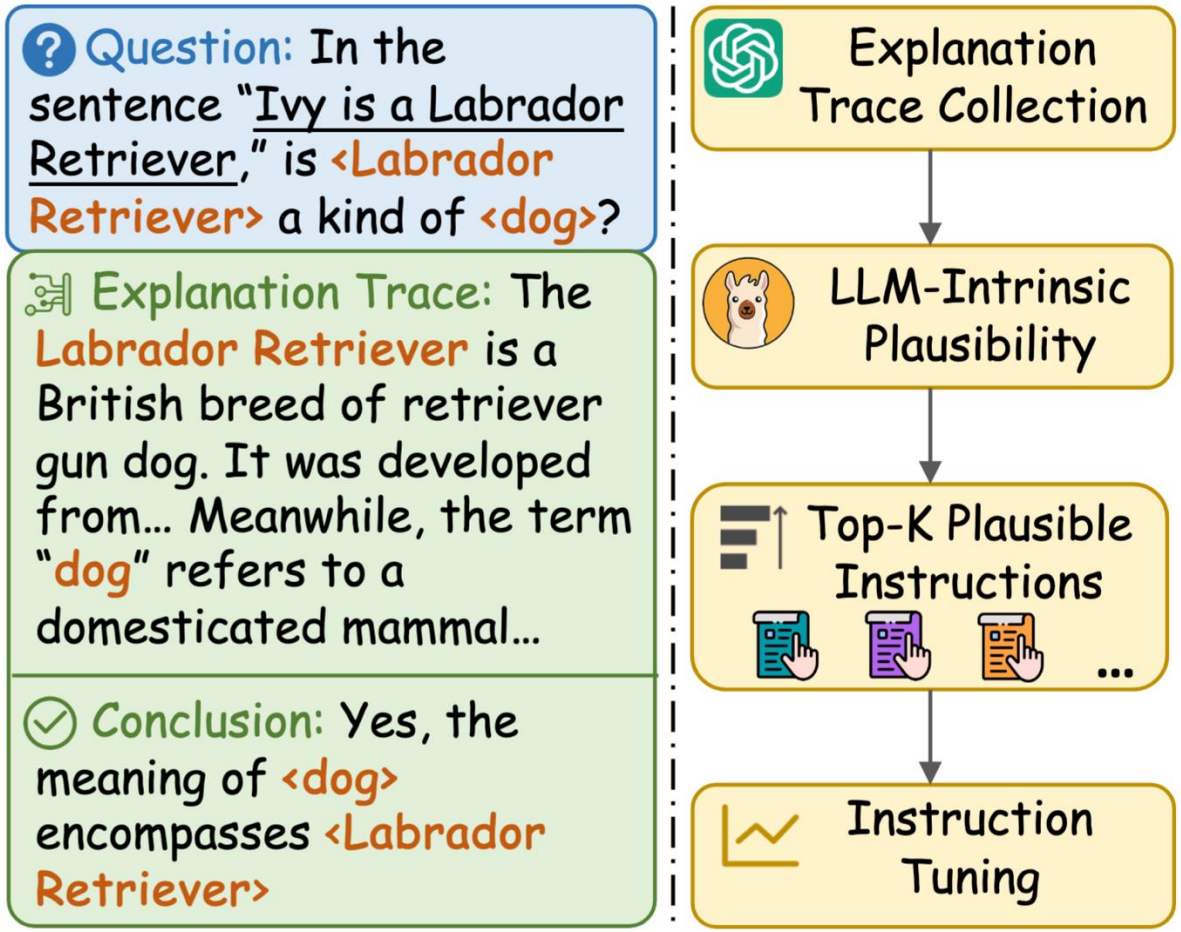
Method

- **Filter Equation** (perplexity reciprocal)

$$Plausibility(i, x, r) = P_\theta(r|i, x)^{\frac{1}{N}},$$

i: instruction; x: example, r: response

- Only 600 examples can work



Result

Data: Test set of AbsPyramid

Our framework can improve perform well on different LLMs

Methods	Backbone	Noun		Verb		Event		All	
		Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1	Acc	Ma-F1
Random	-	50.00	49.56	50.00	49.95	50.00	48.98	50.00	49.50
LLM API (Zero)	GPT 4	79.70	77.34	57.50	54.24	69.70	63.32	68.97	64.97
	GPT 3.5	67.00	62.45	56.30	55.90	65.60	58.23	62.97	58.86
	ChatGPT	74.00	72.27	56.30	55.71	68.20	63.22	66.17	63.73
	ChatGPT (SC)	74.40	72.75	55.50	54.70	68.90	63.49	66.27	63.65
LLM API (10-shot)	GPT 4	70.50	70.49	57.30	56.88	67.20	62.91	65.00	63.43
	GPT 3.5	73.10	71.74	57.20	57.07	66.90	63.79	65.73	64.20
	ChatGPT	76.10	74.60	58.60	58.51	68.90	60.51	67.87	64.54
	ChatGPT (SC)	76.60	75.07	59.10	59.04	68.80	59.56	68.17	64.55
Alpaca (10-shot)	MPT (7B)	43.42	34.71	48.72	37.94	65.33	43.72	52.49	38.79
	Falcon (7B)	60.68	55.07	56.35	56.15	63.92	45.17	60.32	52.13
	Mistral (7B)	76.08	74.10	59.20	58.66	67.66	60.69	67.65	64.49
	Llama2 (7B)	61.96	61.94	55.53	53.19	69.71	60.24	62.40	58.46
	Llama2 (13B)	75.28	72.31	58.97	58.92	66.93	61.73	67.06	64.32
Direct Injection	MPT (7B)	63.87	63.23	53.71	52.37	51.85	51.70	56.47	55.77
	Falcon (7B)	63.48	58.54	55.27	55.16	51.21	51.14	56.66	54.95
	Mistral (7B)	74.90	74.62	59.39	59.11	59.95	59.27	64.74	64.33
	Llama2 (7B)	67.24	66.34	56.66	55.72	55.11	55.11	59.67	59.05
	Llama2 (13B)	75.04	74.09	60.04	59.91	59.26	58.44	64.78	64.15
AbsInstruct	MPT (7B)	71.34	70.89	58.63	58.63	67.52	65.16	65.83	64.89
	Falcon (7B)	66.92	66.45	57.06	56.11	69.03	64.15	64.33	62.24
	Mistral (7B)	80.59	79.85	**60.80**	**60.74**	70.96	66.54	70.78	69.04
	Llama2 (7B)	77.07	75.81	59.44	59.07	**72.72**	**68.00**	69.74	67.63
	Llama2 (13B)	**81.13**	**80.35**	60.58	60.58	71.92	67.24	**71.21**	**69.39**

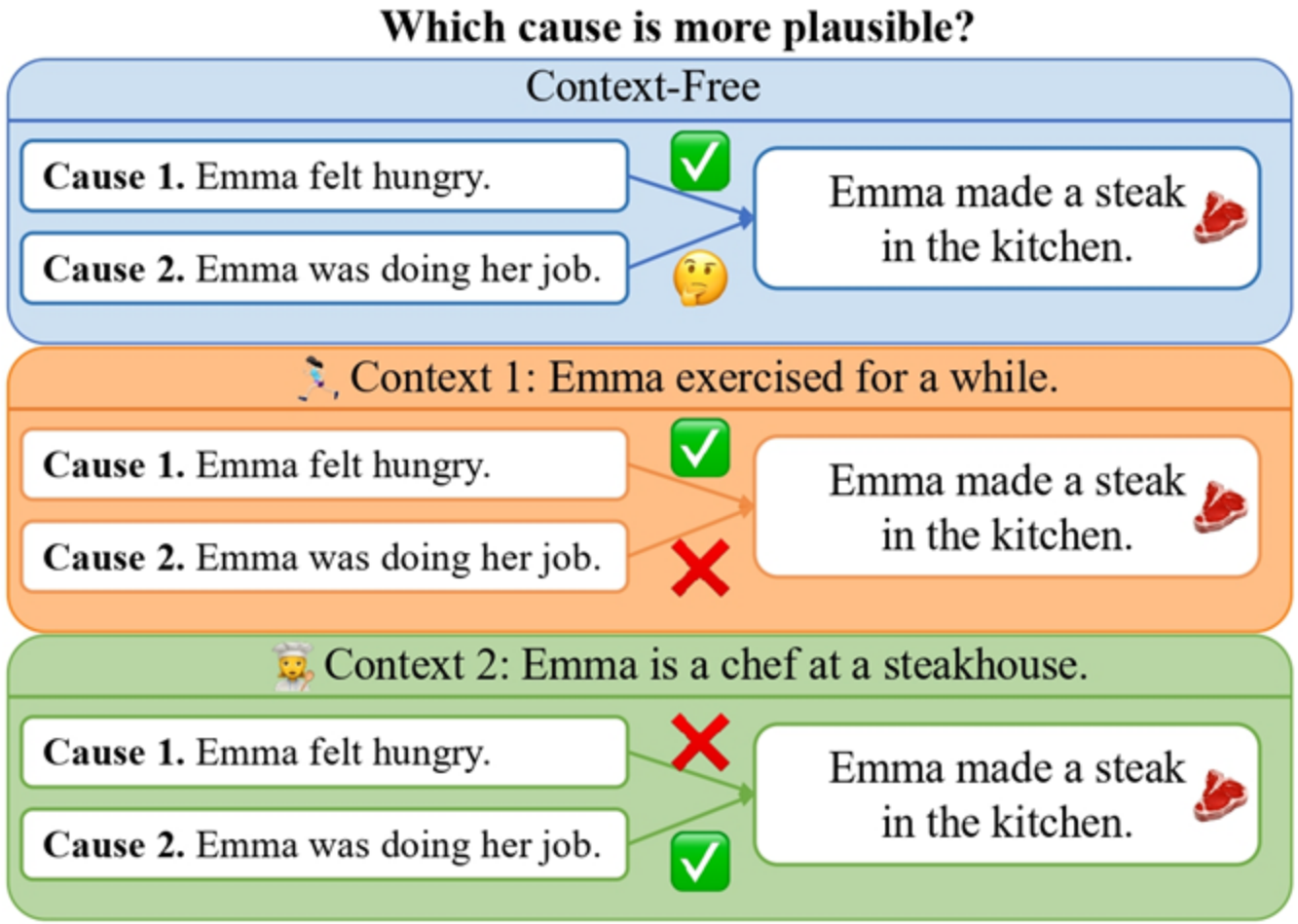
4. COLA: Contextualized Commonsense Causal Reasoning from the Causal Inference Perspective

Zhaowei Wang, Do V. Quyet, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Y. Wong & Simon See

Task Formulation

Contextualized Commonsense Causal Reasoning

In a chain of events E_1, E_2, E_3, …, E_n: detect commonsense causal relation between two events.



Method

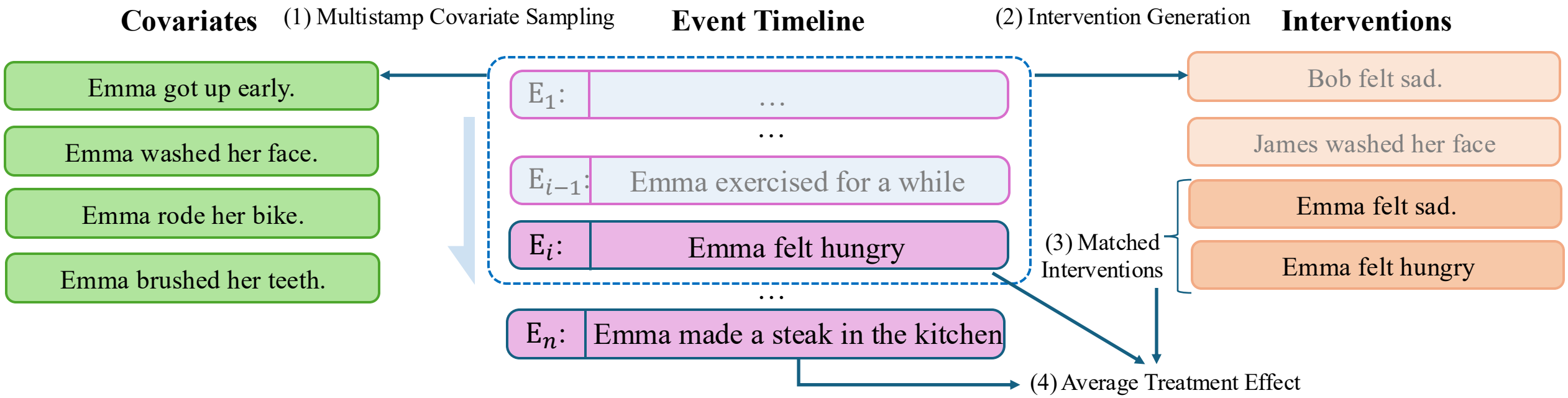
We model the causal relation as Average Treatment Effect (ATE)

$$\Delta = \mathbb{P}\left(\mathsf{E}_i \prec \mathsf{E}_j\right) - \mathbb{P}\left(\neg\mathsf{E}_i \prec \mathsf{E}_j\right),$$

We also use $\mathsf{E}_i \prec \mathsf{E}_j$ to indicate that E_i occurs before E_j for simplicity

So $\mathsf{E}_i \prec \mathsf{E}_j$ is the temporal relation between them.

Method



1. multi-timestamp covariate sampling
2. intervention generation
3. selecting a set of matched interventions
4. computing average treatment effect

Main Results

- COPES dataset, sampled from RocStories[1]
- We test each baseline based on different LMs

Models	Validation				Testing			
	Acc	F1	Ma-F1	Δ_{Acc}	Acc	F1	Ma-F1	Δ_{Acc}
Random	59.47	42.35	55.55	-	58.94	41.10	54.79	-
CLM Perplexity (GPT2)	61.76	45.61	58.06	-	61.47	44.73	57.58	-
CLM Perplexity (GPT2-medium)	60.29	43.51	56.45	-	61.76	45.15	57.90	-
CLM Perplexity (GPT2-large)	62.94	47.28	59.35	-	62.65	46.41	58.87	-
CLM Perplexity (GPT2-XL)	62.65	46.86	59.03	-	62.35	45.99	58.55	-
CLM Perplexity (GPT-J 6b)	63.82	48.54	60.32	-	62.06	45.57	58.22	-
ClozePromptScore (BERT-base)	64.41	49.37	60.97	-	63.53	47.68	59.84	-
ClozePromptScore (BERT-large)	66.47	52.30	63.23	-	62.06	45.57	58.22	-
ClozePromptScore (RoBERTa-base)	59.71	42.68	55.81	-	59.71	42.19	55.63	-
ClozePromptScore (RoBERTa-large)	60.59	43.93	56.77	-	59.12	41.35	54.99	-
ClozePromptScore (DeBERTa-base)	56.76	38.49	52.58	-	58.53	40.51	54.34	-
ClozePromptScore (DeBERTa-large)	56.47	38.08	52.26	-	57.06	38.40	52.72	-
ROCK (BERT-base)	66.18	51.88	62.90	-	65.29	50.21	61.79	-
ROCK (BERT-large)	65.59	51.05	62.26	-	66.47	51.90	63.08	-
ROCK (RoBERTa-base)	61.76	45.61	58.06	-	61.18	44.30	57.25	-
ROCK (RoBERTa-large)	62.94	47.28	59.35	-	65.59	50.63	62.11	-
ROCK (DeBERTa-base)	62.65	46.86	59.03	-	60.59	43.46	56.61	-
ROCK (DeBERTa-large)	64.41	49.37	60.97	-	64.12	48.52	60.49	-
COLA (BERT-base)	67.65	53.97	64.52	↑ 1.47	68.82	55.27	65.67	↑ 3.53
COLA (BERT-large)	70.29	57.74	67.42	↑ 4.70	**70.29**	**57.38**	**67.29**	↑ 3.82
COLA (RoBERTa-base)	69.71	56.90	66.77	↑ **7.95**	66.76	52.32	63.41	↑ 5.58
COLA (RoBERTa-large)	70.59	58.16	67.74	↑ 7.65	70.00	56.96	66.97	↑ 4.41
COLA (DeBERTa-base)	69.71	56.90	66.77	↑ 7.06	**70.29**	**57.38**	**67.29**	↑ **9.70**
COLA (DeBERTa-large)	**71.18**	**59.00**	**68.39**	↑ 6.77	69.41	56.12	66.32	↑ 5.29

[1] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of NAACL*, 2016.