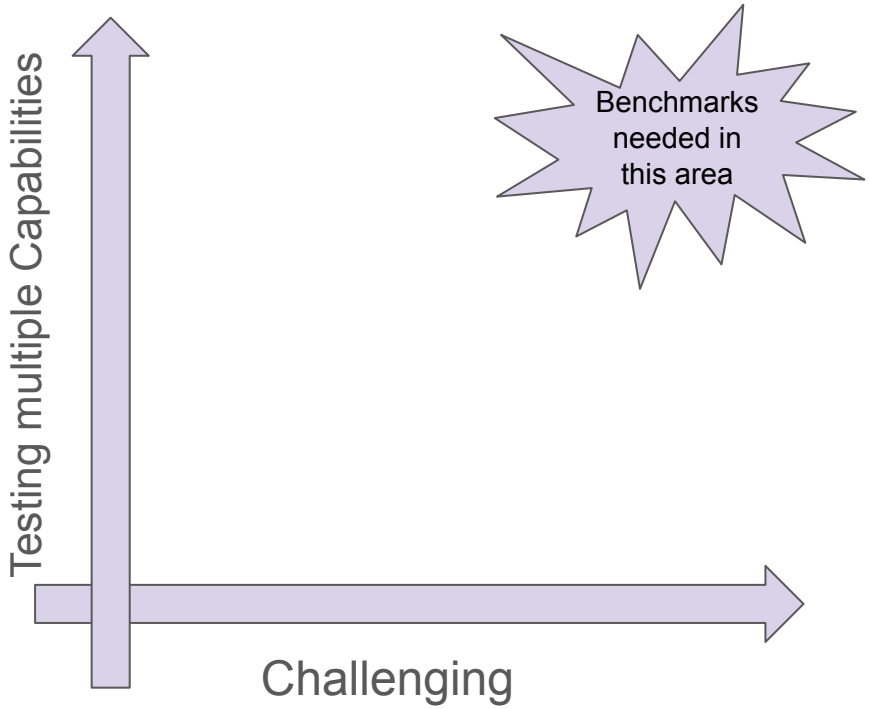
Google DeepMind

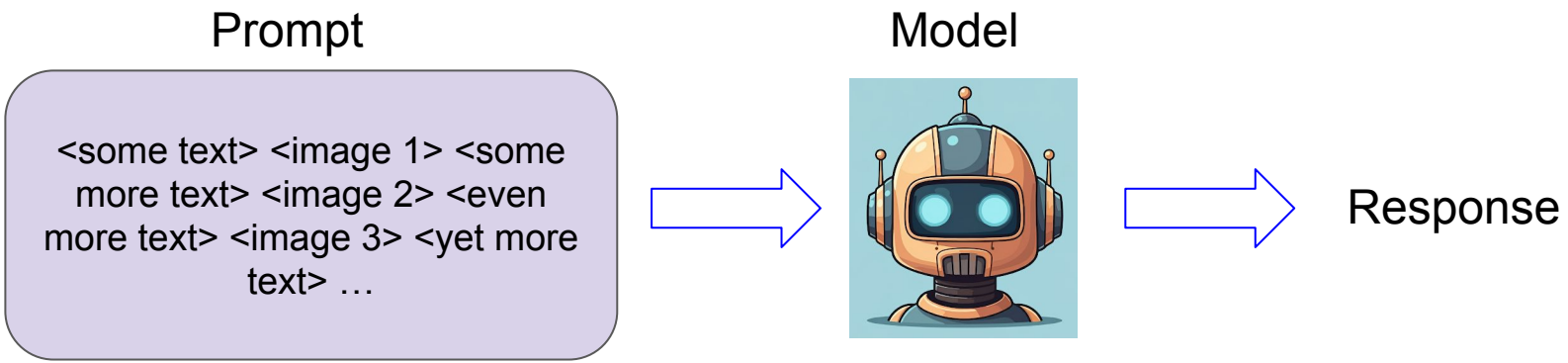
ReMI: A Dataset for Reasoning with Multiple Images

Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Dee Guo, Sreenivas Gollapudi, Ahmed Qureshi

We Need Challenging Benchmarks that Evaluate New Capabilities



Capability: Reasoning over Multiple Images, possibly Interleaved with Text.

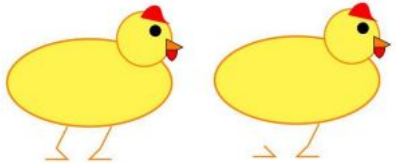


ReMI: A Dataset for Reasoning over Multiple Images

- A combination of 13 diverse tasks (200 examples per task)

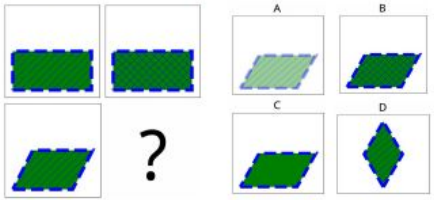
Question: I have some tikz code:
\begin{tikzpicture}
\filldraw[fill=yellow!80, draw=orange, line width=1pt] (0,0) ellipse (1.5cm and 0.8cm);
\filldraw[fill=yellow!80, draw=orange, line width=0.8pt] (1,0.8) circle (0.5cm);\n\n\filldraw[fill=orange] (1.4,0.8) -- (1.7,0.65) -- (1.4, 0.5) -- cycle;\n\n\filldraw[fill=black] (1.2,0.95) circle (0.1cm);

\end{tikzpicture}
It renders to look like:\n<image1>. I instead want it to look like:\n<image2>. You just need to remove one of the lines in the original tikz code to create this new figure. Can you tell me which line to remove? Only quote the line I should remove.



Question: Here are two images. The first image is image A <image1> and the second image is image B <image2>. These images are from Google Maps that depict two different regions around W Glendale Avenue in Glendale, AZ. In which image are there more bus stops along Glendale Avenue. The answer is either 'A', 'B' or 'equal'.

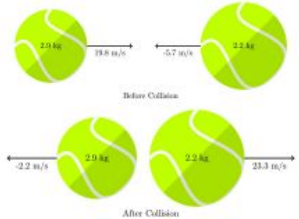
Question: Below is an IQ test <image1>. From the possible options A, B, C, and D shown in the following images in order <image2> <image3> <image4> <image5>, which one logically belongs to the spot of the question mark?



Question: The image below shows the current time <image1>. And the image below shows the departure times for all trains departing today <image2>. Find the destination city for the next scheduled train.

Departure Time	Destination City	Train Name
04:27	Dallas	Amazin Train
17:12	Winsor	Beta Train
17:32	Detroit	Omega Train
18:03	Cancun	Max Train
18:09	Ottawa	Fast Train
20:21	Montreal	Alpha Train

Question: <image1><image2>The images demonstrate the before and after a collision between two balls. Is the momentum conserved in this collision? Answer with 1 if it is conserved, and with 0 if it is not.



...

ReMI: A Dataset for Reasoning over Multiple Images

- A combination of 13 novel tasks,
- Covers multiple domains where multi-image reasoning arise:
  - Math (algebra, geometry, calculus),
  - Code,
  - Spatial reasoning,
  - Temporal reasoning,
  - Logic,
  - Physics,
  - …

ReMI: A Dataset for Reasoning over Multiple Images

- A combination of 13 novel tasks,
- Covers multiple domains where multi-image reasoning arise
- Covers many distinctive features that arise in multi-image reasoning:

Sequence vs set

Whether the provided images have to be consumed sequentially or as a set.

ReMI: A Dataset for Reasoning over Multiple Images

- A combination of 13 novel tasks,
- Covers multiple domains where multi-image reasoning arise
- Covers many distinctive features that arise in multi-image reasoning:

Sequence vs set

Whether the provided images have to be consumed sequentially or as a set.

Same/different concept

Whether the provided images all exhibit the same concept or different concepts.

ReMI: A Dataset for Reasoning over Multiple Images

- A combination of 13 novel tasks,
- Covers multiple domains where multi-image reasoning arise
- Covers many distinctive features that arise in multi-image reasoning:

Sequence vs set

Whether the provided images have to be consumed sequentially or as a set.

Same/different concept

Whether the provided images all exhibit the same concept or different concepts.

Interleaved

Whether the images are interleaved with text or all provided at the beginning of the prompt.

ReMI: A Dataset for Reasoning over Multiple Images

- A combination of 13 novel tasks,
- Covers multiple domains where multi-image reasoning arise
- Covers many distinctive features that arise in multi-image reasoning:

Sequence vs set

Whether the provided images have to be consumed sequentially or as a set.

Same/different concept

Whether the provided images all exhibit the same concept or different concepts.

Interleaved

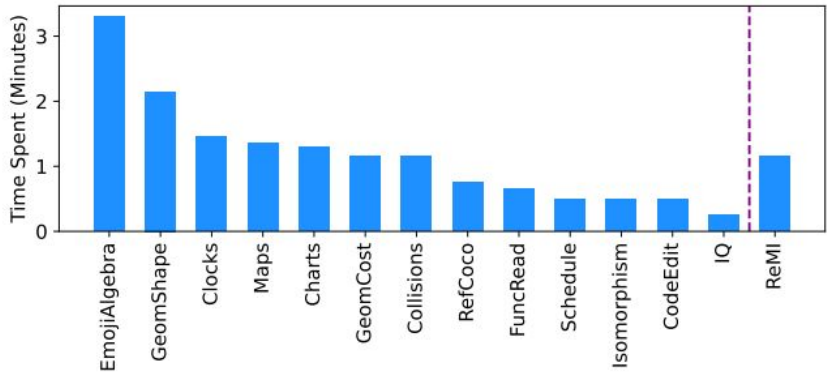
Whether the images are interleaved with text or all provided at the beginning of the prompt.

Number of images per task

Each of our problems contains between 2 to 6 images.

ReMI: A Dataset for Reasoning over Multiple Images

- A combination of 13 novel tasks,
- Covers multiple domains where multi-image reasoning arise
- Covers many distinctive features that arise in multi-image reasoning
- On average, Humans require different amounts of time to solve each of the 13 tasks in ReMI:

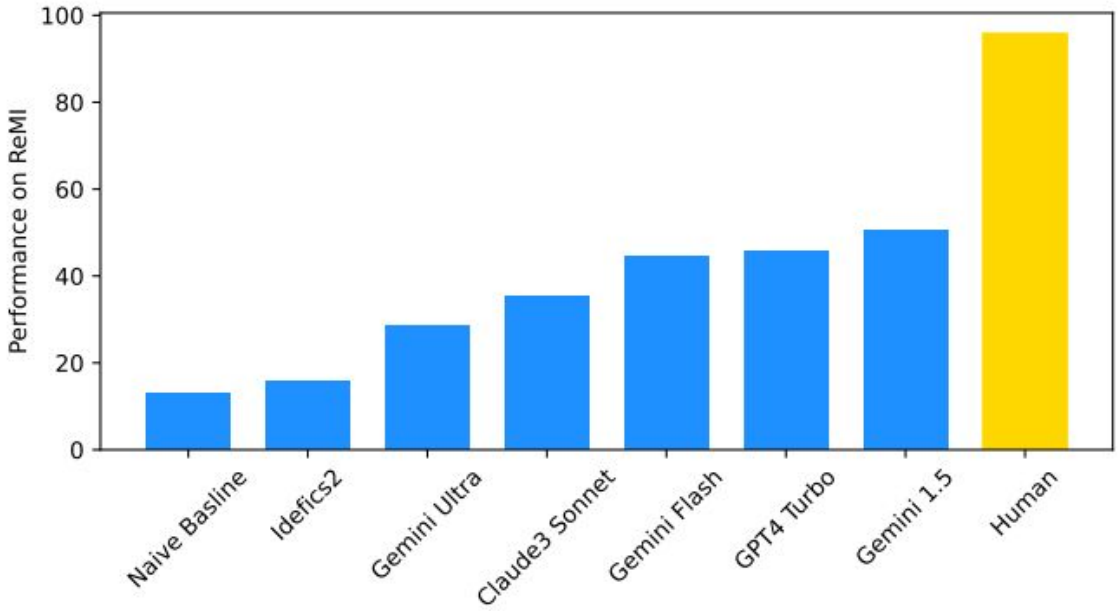


ReMI: A Dataset for Reasoning over Multiple Images

- A combination of 13 novel tasks,
- Covers multiple domains where multi-image reasoning arise
- Covers many distinctive features that arise in multi-image reasoning
- ReMI contains questions with a variety of lengths
- On average, Humans require different amounts of time to solve each of the 13 tasks in ReMI:

Overall, ReMI is quite diverse, covering many multi-image domains and properties, with problems of varying complexity.

Main Result: SoTA models are far below human performance

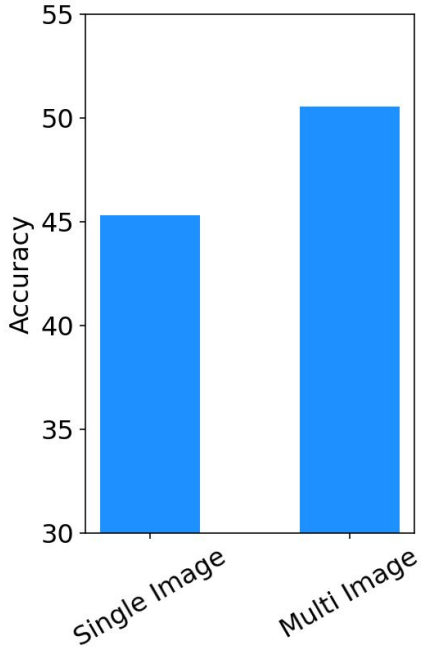


Task-Level Results: Different models perform better on different tasks

Task Name	Naive Baseline	Idefics2	Claude3 Sonnet	Gemini Ultra	Gemini Flash	Gemini 1.5	GPT4 Turbo	Human
EmojiAlgebra	0.0	1.0	28.0	2.5	15.0	44.5	**57.5**	100.0
FuncRead	5.5	11.0	24.0	15.0	36.0	**40.0**	26.0	100.0
GeomShapes	0.0	14.0	17.5	14.5	34.0	**51.5**	32.5	100.0
GeomCost	0.0	2.0	58.5	47.0	75.0	**81.5**	70.5	90.0
Collisions	30.8	31.5	51.5	36.5	56.5	50.5	**62.0**	100.0
Clocks	2.0	3.0	**5.0**	4.0	4.0	2.5	4.0	80.0
Schedule	0.0	21.5	36.0	33.0	43.0	40.5	**49.5**	90.0
Charts	2.5	1.0	40.0	30.0	53.0	**54.0**	44.0	95.0
CodeEdit	14.9	12.5	20.0	24.5	**46.0**	41.0	42.0	95.0
Isomorphism	50.0	35.5	57.0	65.0	67.0	**72.0**	71.5	100.0
Maps	28.0	38.0	39.5	39.0	**47.0**	**47.0**	36.5	100.0
RefCOCO	12.0	14.5	30.0	31.0	49.0	**56.0**	37.5	95.0
IQ	25.0	19.0	50.5	30.0	53.0	**76.0**	62.5	100.0
ReMI	13.1	15.7	35.2	28.6	44.5	**50.5**	45.8	95.8

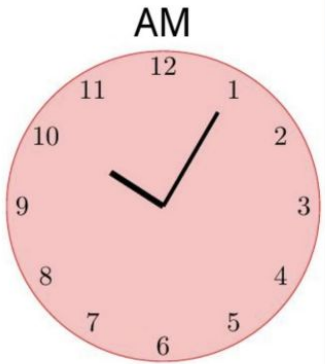
Combining images into a single images only makes things worse

This is especially true for the interleaved tasks.



Failure Analysis

- We do a thorough failure analysis which reveals several shortcomings and areas for improvement.
- One interesting one in particular is that models cannot correctly read time from analog clocks:



Model	Time Read
Gemini Ultra	1:00 AM
Gemini Flash	9:15 AM
Gemini 1.5	1:50 AM
Claude3 Sonnet	3:25 PM
GPT4 Turbo	5:00 AM

Conclusion

- We developed ReMI, a dataset dedicated to reasoning with multiple images.
- ReMI contains 13 diverse tasks and covers a wide range of different domains as well as different characteristics of multi-image reasoning problems.
- The performance of SoTA models is still far behind human performance.
- Dataset available on HuggingFace:
  - https://huggingface.co/datasets/mehrankazemi/ReMI