NEURAL INFORMATION PROCESSING SYSTEMS

SciFIBench: Benchmarking Large Multimodal Models for Scientific Figure Interpretation

Jonathan Roberts[1], Kai Han[2], Neil Houlsby[3], Samuel Albanie

[1]University of Cambridge

[2]The University of Hong Kong

[3]Google DeepMind

UNIVERSITY OF CAMBRIDGE

香 港 大 學 THE UNIVERSITY OF HONG KONG

Google DeepMind

Motivation

- The capabilities of Large Multimodal Models (LMMs) have been demonstrated in many domains
- LMMs have the potential to benefit the scientific domain
- A tool to assist different stages of the scientific process
- Understanding figures is a key component of scientific research
- The capacity of LMMs to understand scientific figures is not well-known

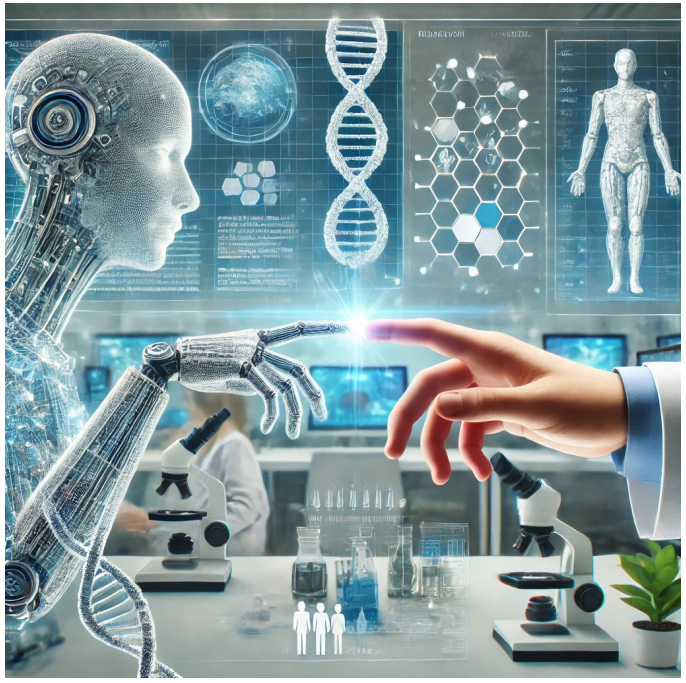


Image generated using DALL-E

OpenAI. (2024). DALL-E Image Generation Model (Version 3). OpenAI. Available from https://openai.com/dall-e

Curation

SciFIBench (Scientific Figure Interpretation Benchmark)

Source datasets: SciCap [1] and ArXivCap [2]

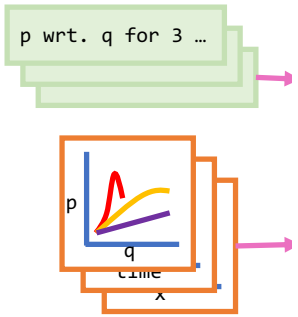
[1] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. arXiv preprint arXiv:2403.00231, 2024.

[2] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624, 2021.

Curation

SciFIBench (Scientific Figure Interpretation Benchmark)

Curation (Figure → Caption)



(1) arXiv figures and captions

Source datasets: SciCap [1] and ArXivCap [2]

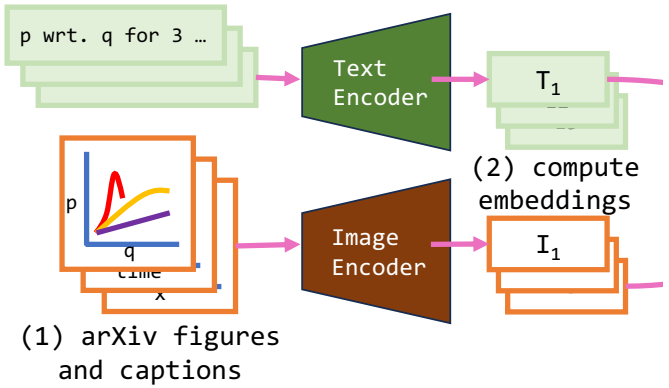
[1] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. arXiv preprint arXiv:2403.00231, 2024.

[2] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624, 2021.

Curation

SciFIBench (Scientific Figure Interpretation Benchmark)

Curation (Figure → Caption)



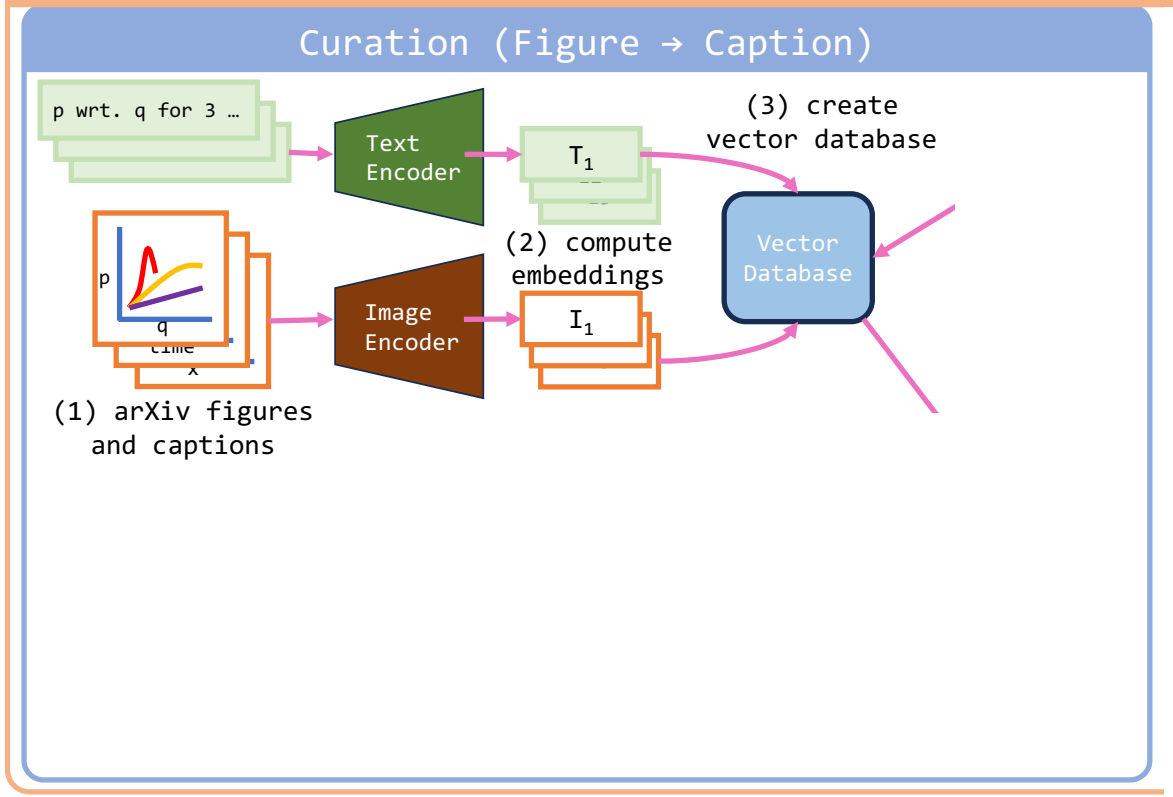
Source datasets: SciCap [1] and ArXivCap [2]

[1] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. arXiv preprint arXiv:2403.00231, 2024.

[2] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624, 2021.

Curation

SciFIBench (Scientific Figure Interpretation Benchmark)



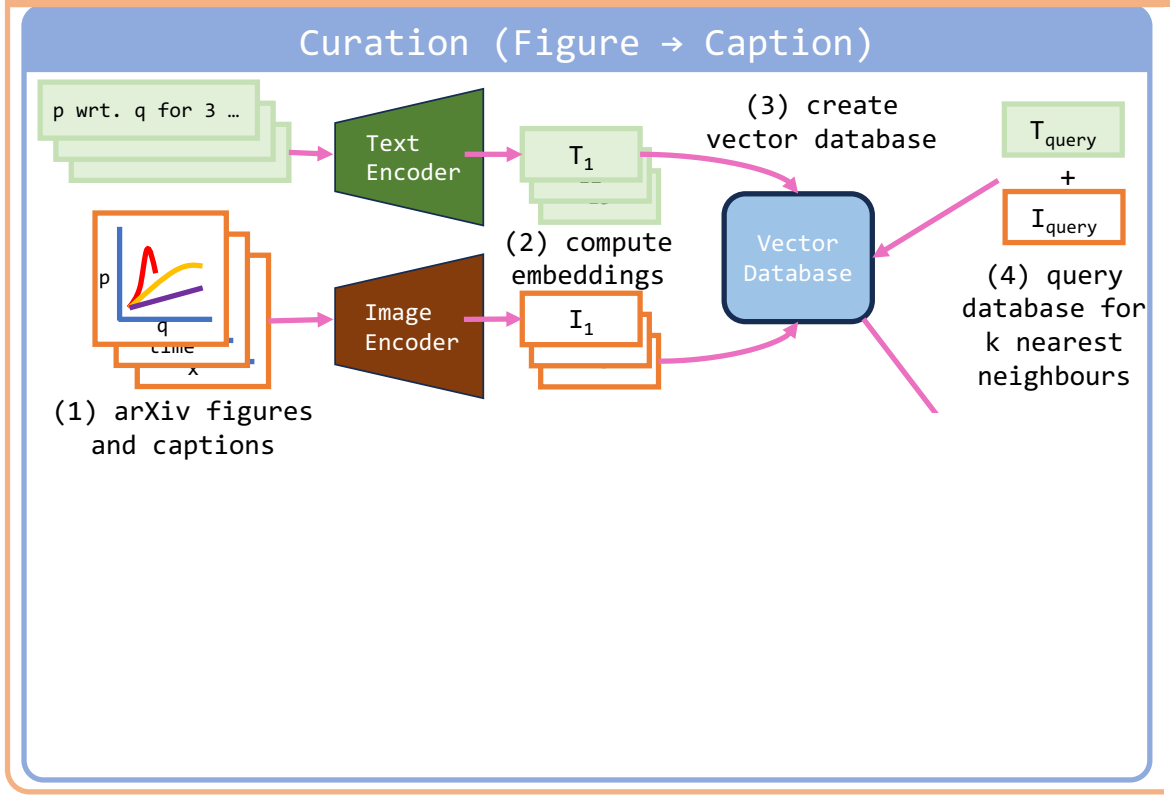
Source datasets: SciCap [1] and ArXivCap [2]

[1] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. arXiv preprint arXiv:2403.00231, 2024.

[2] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624, 2021.

Curation

SciFIBench (Scientific Figure Interpretation Benchmark)



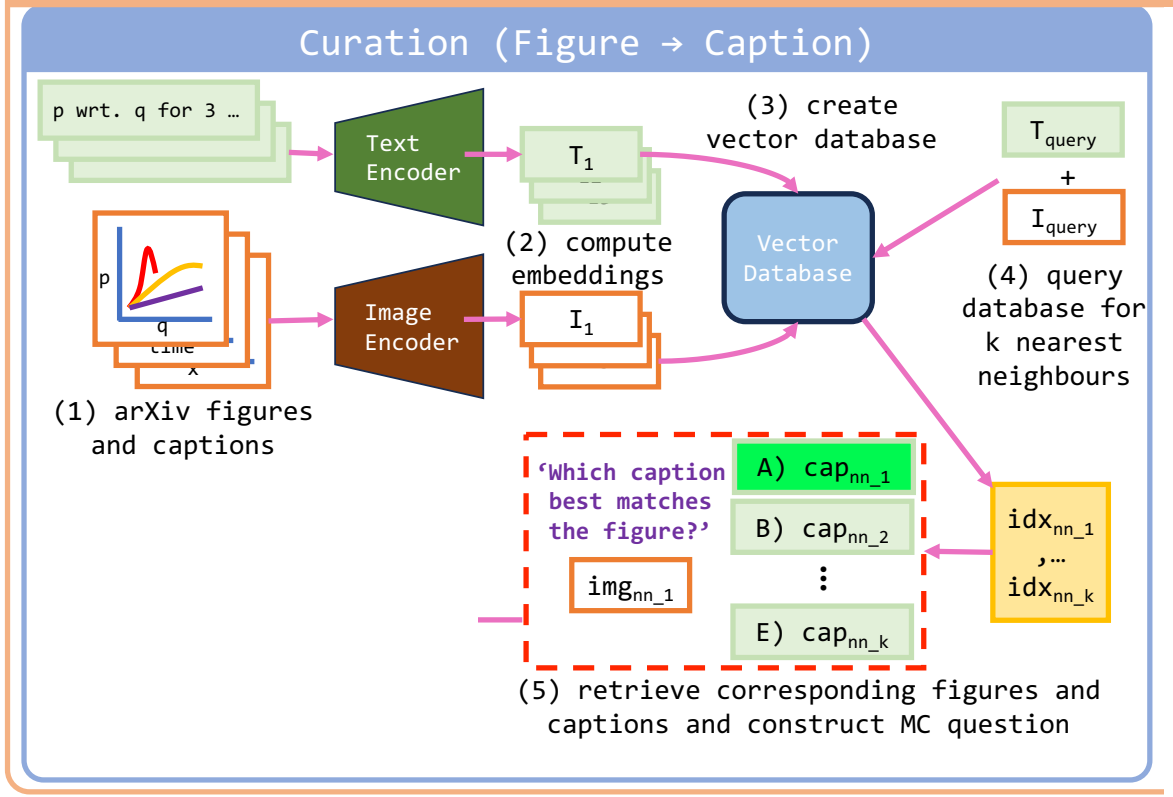
Source datasets: SciCap [1] and ArXivCap [2]

[1] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. arXiv preprint arXiv:2403.00231, 2024.

[2] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624, 2021.

Curation

SciFIBench (Scientific Figure Interpretation Benchmark)



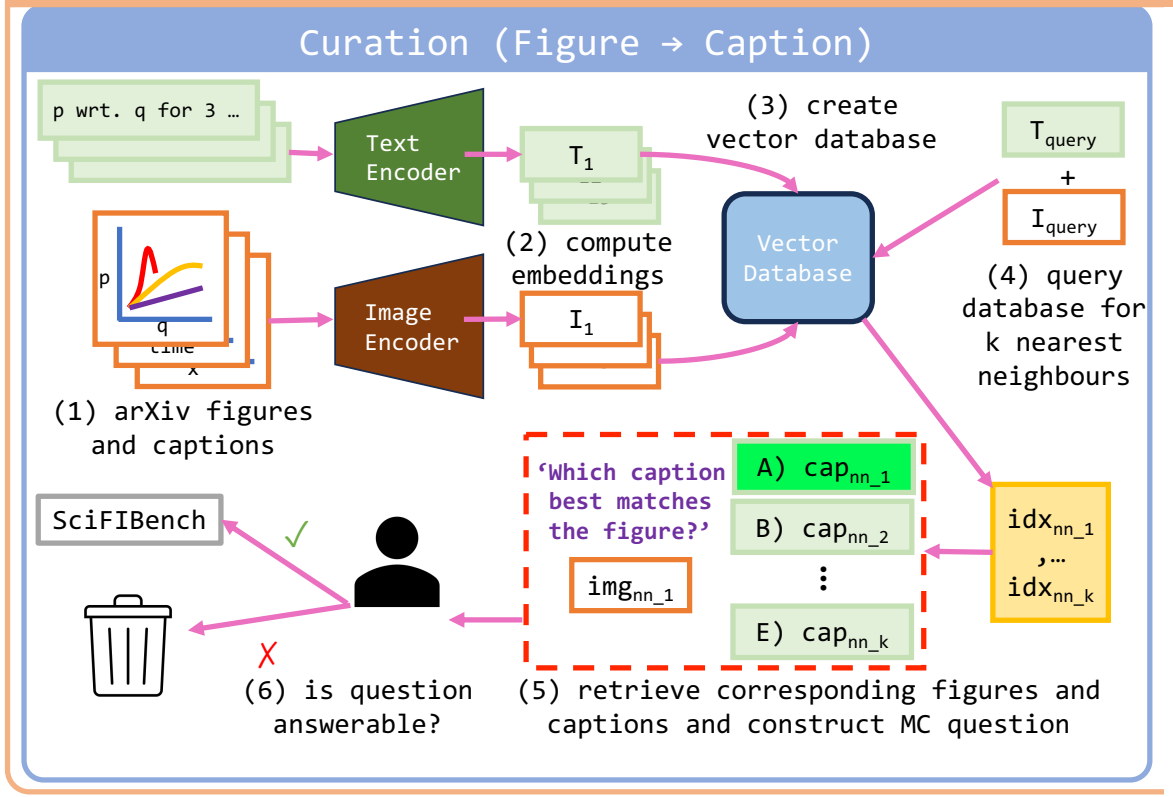
Source datasets: SciCap [1] and ArXivCap [2]

[1] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. arXiv preprint arXiv:2403.00231, 2024.

[2] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624, 2021.

Curation

SciFIBench (Scientific Figure Interpretation Benchmark)

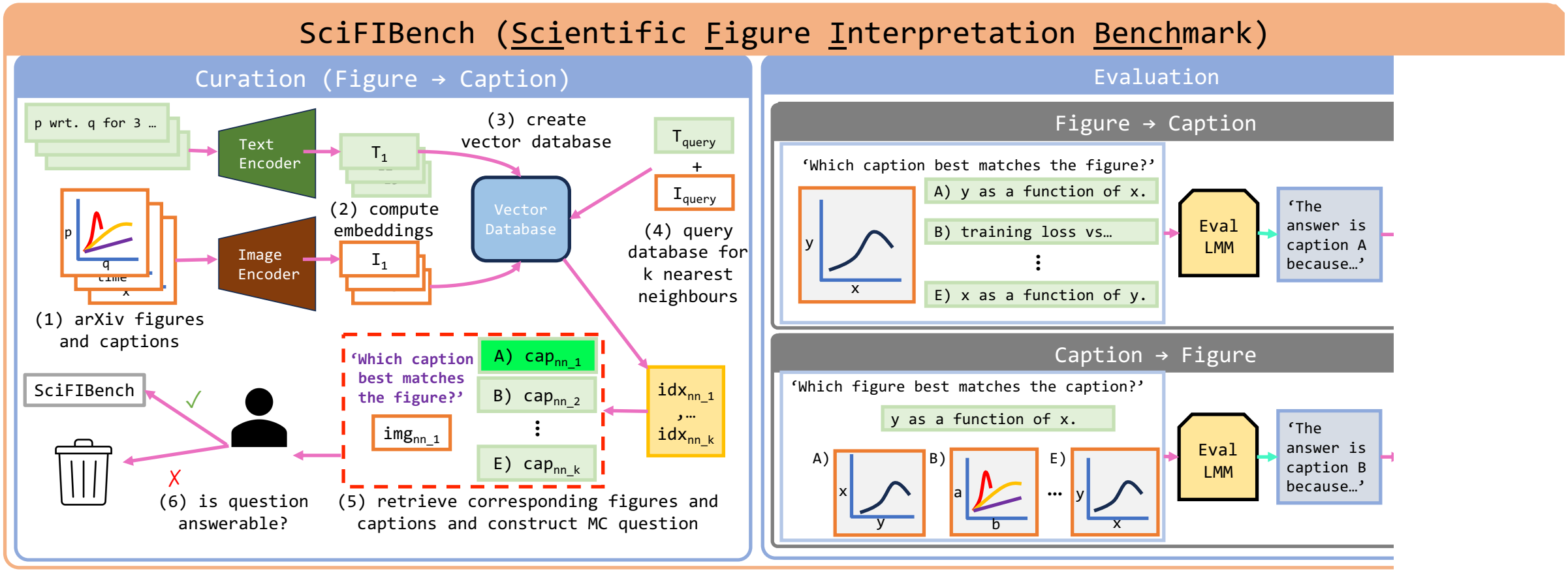


Source datasets: SciCap [1] and ArXivCap [2]

[1] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. arXiv preprint arXiv:2403.00231, 2024.

[2] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624, 2021.

Curation

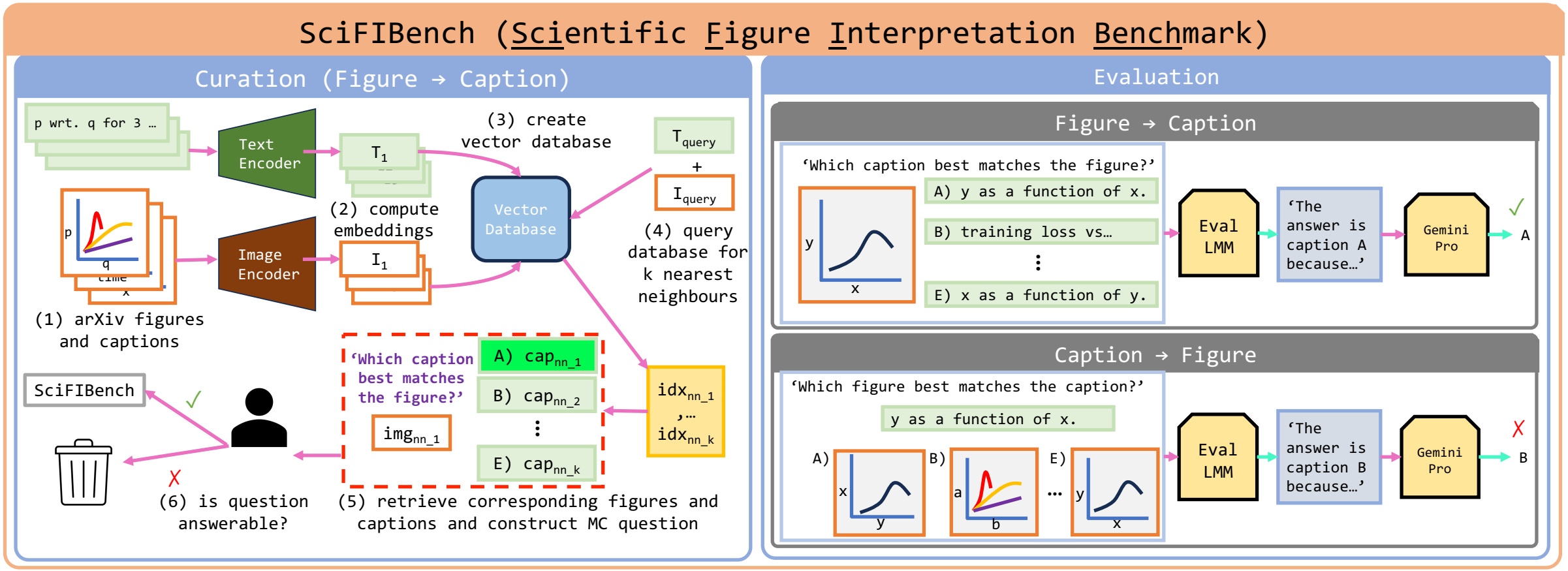


Source datasets: SciCap [1] and ArXivCap [2]

[1] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. arXiv preprint arXiv:2403.00231, 2024.

[2] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624, 2021.

Curation



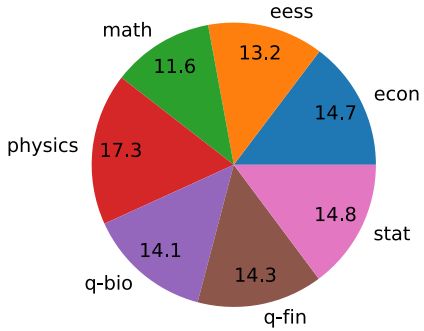
Source datasets: SciCap [1] and ArXivCap [2]

[1] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. arXiv preprint arXiv:2403.00231, 2024.

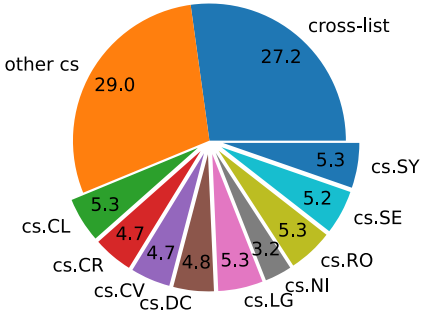
[2] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624, 2021.

Statistics & Examples

- 2 tasks
- 2000 questions
- 2 subsets: General and Computer Science



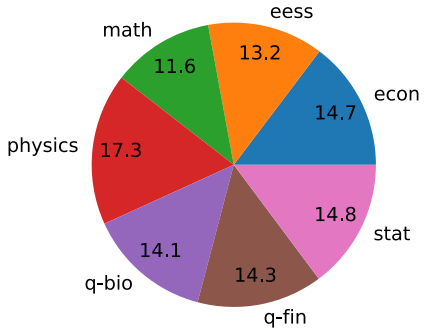
General



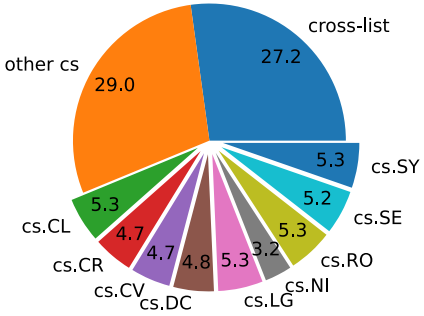
Computer Science

Statistics & Examples

- 2 tasks
- 2000 questions
- 2 subsets: General and Computer Science

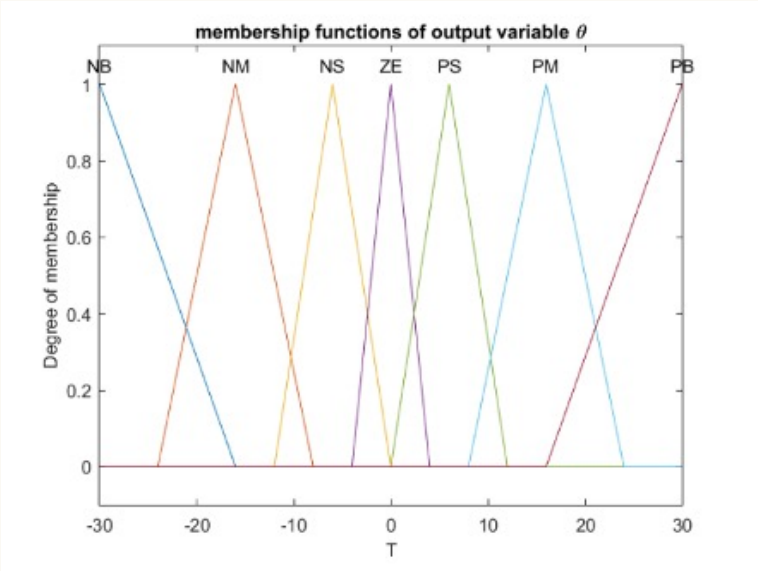


General



Computer Science

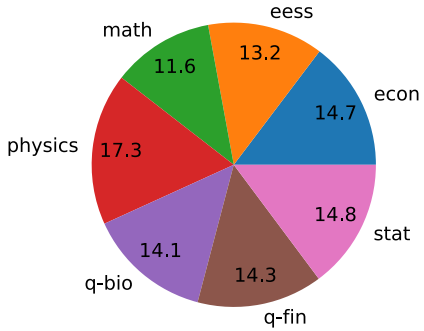
Figure -> Caption



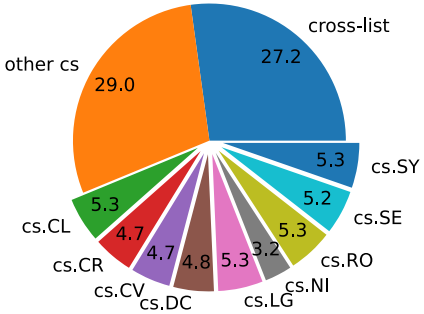
A) illustration of the designed membership functions of α .
B) illustration of the designed membership functions of β .
C) illustration of the designed membership functions of γ .
D) illustration of the designed membership functions of θ .
E) illustration of the designed membership functions of x .

Statistics & Examples

- 2 tasks
- 2000 questions
- 2 subsets: General and Computer Science

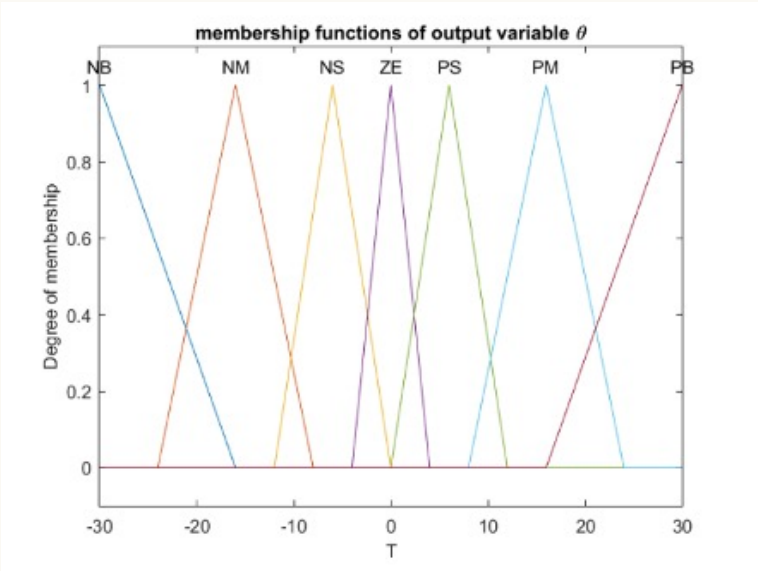


General



Computer Science

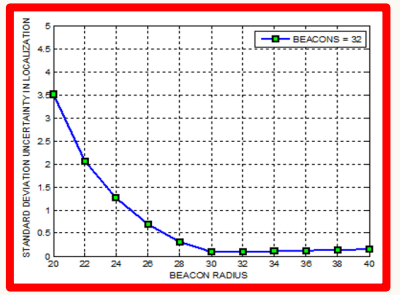
Figure -> Caption



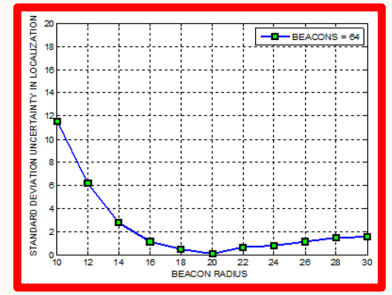
A) illustration of the designed membership functions of α .
B) illustration of the designed membership functions of β .
C) illustration of the designed membership functions of γ .
D) illustration of the designed membership functions of θ .
E) illustration of the designed membership functions of x .

Caption -> Figure

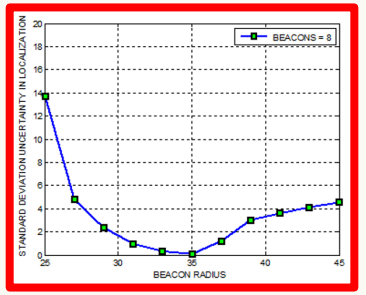
standard deviation uncertainty in localization v/s beacon radius for number of beacons = 16 .



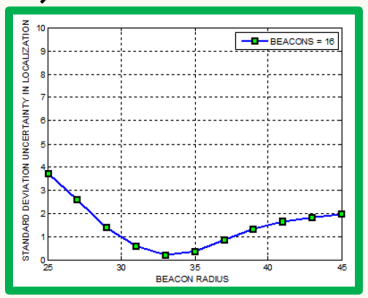
A)



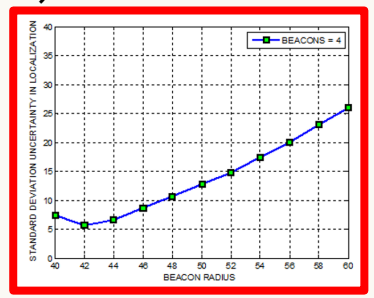
B)



C)



D)



E)

Experiments

Model	CS		General		Overall	
	Fig.→Cap.	Cap.→Fig.	Fig.→Cap.	Cap.→Fig.	Fig.→Cap.	Cap.→Fig.
Closed-source LMMs						
GPT-4V [5]	69.4	58.4	-	-	-	-
GPT-4 Turbo [52]	68.0	60.6	62.8	55.2	65.4	57.9
GPT-4o [31]	**75.4**	**72.2**	**72.2**	58.6	**73.8**	65.4
Gemini Pro Vision [7]	56.0	52.4	50.6	39.6	53.3	46.0
Gemini 1.5 Pro [32]	74.0	**76.0**	65.2	56.2	69.6	**66.1**
Gemini 1.5 Flash [32]	74.4	69.6	65.8	**62.4**	70.1	**66.1**
Claude 3 Haiku [53]	52.6	43.8	52.6	33.0	52.6	38.4
Claude 3 Sonnet [53]	53.4	58.4	53.6	55.0	53.5	56.7
Claude 3 Opus [53]	59.8	49.2	50.8	47.4	55.3	48.3
Open-source LMMs						
IDEFICS-9b-Instruct [54]	20.6	20.2	17.6	12.6	19.1	16.4
IDEFICS-80b-Instruct [54]	**20.6**	**24.2**	18.4	**20.6**	19.5	**22.4**
Qwen-VL-Chat [6]	28.0	16.0	17.0	19.2	**22.5**	17.6
Emu2 [55]	20.8	-	19.6	-	20.2	-
TransCore-M [56]	**51.0**	-	**27.4**	-	**39.2**	-
InternLM-XComposer-7b [57]	34.0	-	21.6	-	27.8	-
InternLM-XComposer2-7b [57]	28.0	-	23.8	-	25.9	-
CogVLM-Chat [58]	40.8	-	24.0	-	32.4	-
OmniLMM-3b [59]	35.8	-	24.8	-	30.3	-
OmniLMM-12b [59]	34.2	-	27.2	-	30.7	-
Yi-VL-6b [60]	41.4	-	27.0	-	34.2	-
Yi-VL-34b [60]	32.6	-	21.4	-	27.0	-
InstructBLIP-FlanT5-xl [61]	35.8	-	19.0	-	27.4	-
InstructBLIP-FlanT5-xxl [61]	**36.2**	-	**26.8**	-	**31.5**	-
InstructBLIP-Vicuna-7b [61]	21.0	-	12.8	-	16.9	-
InstructBLIP-Vicuna-13b [61]	22.2	-	15.6	-	18.9	-
Monkey-Chat [62]	27.2	-	18.2	-	22.7	-
LLaVA-1.5-7b [63]	32.8	-	22.8	-	27.8	-
LLaVA-1.5-13b [63]	25.0	-	20.2	-	22.6	-
Text-only input						
Gemini-Pro 1.5 Flash [32]	**48.0**	**39.2**	**51.0**	**35.8**	**49.5**	**37.5**
VLMs						
CLIP ViT-H-14-378-quickgelu [50]	41.8	42.6	30.6	30.0	36.2	36.3
MetaCLIP ViT-H-14-quickgelu [65]	36.6	35.4	24.2	25.2	30.4	30.3
Google Multimodal Embedding [66]	**47.6**	**54.4**	**28.2**	**28.4**	**37.9**	**41.4**
Human (25 questions per task*)						
Human ($\mu \pm \sigma$)	**86.4**±8.24	**78.4**±8.24	-	-	-	-
GPT-4o	72.0	76.0	-	-	-	-
Gemini-Pro 1.5	84.0	72.0	-	-	-	-
CLIP ViT-H-14-378-quickgelu	48.0	56.0	-	-	-	-
TransCore-M	36.0	-	-	-	-	-

Experiments

Model	CS Fig.→Cap.	CS Cap.→Fig.	General Fig.→Cap.	General Cap.→Fig.	Overall Fig.→Cap.	Overall Cap.→Fig.
Closed-source LMMs						
GPT-4V [5]	69.4	58.4	-	-	-	-
GPT-4 Turbo [52]	68.0	60.6	62.8	55.2	65.4	57.9
GPT-4o [31]	**75.4**	**72.2**	**72.2**	58.6	**73.8**	65.4
Gemini Pro Vision [7]	56.0	52.4	50.6	39.6	53.3	46.0
Gemini 1.5 Pro [32]	74.0	**76.0**	65.2	56.2	69.6	**66.1**
Gemini 1.5 Flash [32]	74.4	69.6	65.8	**62.4**	70.1	**66.1**
Claude 3 Haiku [53]	52.6	43.8	52.6	33.0	52.6	38.4
Claude 3 Sonnet [53]	53.4	58.4	53.6	55.0	53.5	56.7
Claude 3 Opus [53]	59.8	49.2	50.8	47.4	55.3	48.3
Open-source LMMs						
IDEFICS-9b-Instruct [54]	20.6	20.2	17.6	12.6	19.1	16.4
IDEFICS-80b-Instruct [54]	**20.6**	**24.2**	18.4	**20.6**	19.5	**22.4**
Qwen-VL-Chat [6]	28.0	16.0	17.0	19.2	**22.5**	17.6
Emu2 [55]	20.8	-	19.6	-	20.2	-
TransCore-M [56]	**51.0**	-	**27.4**	-	**39.2**	-
InternLM-XComposer-7b [57]	34.0	-	21.6	-	27.8	-
InternLM-XComposer2-7b [57]	28.0	-	23.8	-	25.9	-
CogVLM-Chat [58]	40.8	-	24.0	-	32.4	-
OmniLMM-3b [59]	35.8	-	24.8	-	30.3	-
OmniLMM-12b [59]	34.2	-	27.2	-	30.7	-
Yi-VL-6b [60]	41.4	-	27.0	-	34.2	-
Yi-VL-34b [60]	32.6	-	21.4	-	27.0	-
InstructBLIP-FlanT5-xl [61]	35.8	-	19.0	-	27.4	-
InstructBLIP-FlanT5-xxl [61]	**36.2**	-	**26.8**	-	**31.5**	-
InstructBLIP-Vicuna-7b [61]	21.0	-	12.8	-	16.9	-
InstructBLIP-Vicuna-13b [61]	22.2	-	15.6	-	18.9	-
Monkey-Chat [62]	27.2	-	18.2	-	22.7	-
LLaVA-1.5-7b [63]	32.8	-	22.8	-	27.8	-
LLaVA-1.5-13b [63]	25.0	-	20.2	-	22.6	-
Text-only input						
Gemini-Pro 1.5 Flash [32]	**48.0**	**39.2**	**51.0**	**35.8**	**49.5**	**37.5**
VLMs						
CLIP ViT-H-14-378-quickgelu [50]	41.8	42.6	30.6	30.0	36.2	36.3
MetaCLIP ViT-H-14-quickgelu [65]	36.6	35.4	24.2	25.2	30.4	30.3
Google Multimodal Embedding [66]	**47.6**	**54.4**	**28.2**	**28.4**	**37.9**	**41.4**
Human (25 questions per task*)						
Human ($\mu \pm \sigma$)	**86.4**±8.24	**78.4**±8.24	-	-	-	-
GPT-4o	72.0	76.0	-	-	-	-
Gemini-Pro 1.5	84.0	72.0	-	-	-	-
CLIP ViT-H-14-378-quickgelu	48.0	56.0	-	-	-	-
TransCore-M	36.0	-	-	-	-	-

- SciFIBench is a challenging benchmark

Experiments

Model	CS Fig.→Cap.	CS Cap.→Fig.	General Fig.→Cap.	General Cap.→Fig.	Overall Fig.→Cap.	Overall Cap.→Fig.
Closed-source LMMs						
GPT-4V [5]	69.4	58.4	-	-	-	-
GPT-4 Turbo [52]	68.0	60.6	62.8	55.2	65.4	57.9
GPT-4o [31]	**75.4**	**72.2**	**72.2**	58.6	**73.8**	65.4
Gemini Pro Vision [7]	56.0	52.4	50.6	39.6	53.3	46.0
Gemini 1.5 Pro [32]	74.0	**76.0**	65.2	56.2	69.6	**66.1**
Gemini 1.5 Flash [32]	74.4	69.6	65.8	**62.4**	70.1	**66.1**
Claude 3 Haiku [53]	52.6	43.8	52.6	33.0	52.6	38.4
Claude 3 Sonnet [53]	53.4	58.4	53.6	55.0	53.5	56.7
Claude 3 Opus [53]	59.8	49.2	50.8	47.4	55.3	48.3
Open-source LMMs						
IDEFICS-9b-Instruct [54]	20.6	20.2	17.6	12.6	19.1	16.4
IDEFICS-80b-Instruct [54]	**20.6**	**24.2**	18.4	**20.6**	19.5	**22.4**
Qwen-VL-Chat [6]	28.0	16.0	17.0	19.2	**22.5**	17.6
Emu2 [55]	20.8	-	19.6	-	20.2	-
TransCore-M [56]	**51.0**	-	**27.4**	-	**39.2**	-
InternLM-XComposer-7b [57]	34.0	-	21.6	-	27.8	-
InternLM-XComposer2-7b [57]	28.0	-	23.8	-	25.9	-
CogVLM-Chat [58]	40.8	-	24.0	-	32.4	-
OmniLMM-3b [59]	35.8	-	24.8	-	30.3	-
OmniLMM-12b [59]	34.2	-	27.2	-	30.7	-
Yi-VL-6b [60]	41.4	-	27.0	-	34.2	-
Yi-VL-34b [60]	32.6	-	21.4	-	27.0	-
InstructBLIP-FlanT5-xl [61]	35.8	-	19.0	-	27.4	-
InstructBLIP-FlanT5-xxl [61]	**36.2**	-	**26.8**	-	**31.5**	-
InstructBLIP-Vicuna-7b [61]	21.0	-	12.8	-	16.9	-
InstructBLIP-Vicuna-13b [61]	22.2	-	15.6	-	18.9	-
Monkey-Chat [62]	27.2	-	18.2	-	22.7	-
LLaVA-1.5-7b [63]	32.8	-	22.8	-	27.8	-
LLaVA-1.5-13b [63]	25.0	-	20.2	-	22.6	-
Text-only input						
Gemini-Pro 1.5 Flash [32]	**48.0**	**39.2**	**51.0**	**35.8**	**49.5**	**37.5**
VLMs						
CLIP ViT-H-14-378-quickgelu [50]	41.8	42.6	30.6	30.0	36.2	36.3
MetaCLIP ViT-H-14-quickgelu [65]	36.6	35.4	24.2	25.2	30.4	30.3
Google Multimodal Embedding [66]	**47.6**	**54.4**	**28.2**	**28.4**	**37.9**	**41.4**
Human (25 questions per task*)						
Human ($\mu \pm \sigma$)	**86.4**±8.24	**78.4**±8.24	-	-	-	-
GPT-4o	72.0	76.0	-	-	-	-
Gemini-Pro 1.5	84.0	72.0	-	-	-	-
CLIP ViT-H-14-378-quickgelu	48.0	56.0	-	-	-	-
TransCore-M	36.0	-	-	-	-	-

- SciFIBench is a challenging benchmark
- Closed-source models perform better

Experiments

Model	CS		General		Overall	
	Fig.→Cap.	Cap.→Fig.	Fig.→Cap.	Cap.→Fig.	Fig.→Cap.	Cap.→Fig.
Closed-source LMMs						
GPT-4V [5]	69.4	58.4	-	-	-	-
GPT-4 Turbo [52]	68.0	60.6	62.8	55.2	65.4	57.9
GPT-4o [31]	**75.4**	**72.2**	**72.2**	58.6	**73.8**	65.4
Gemini Pro Vision [7]	56.0	52.4	50.6	39.6	53.3	46.0
Gemini 1.5 Pro [32]	74.0	**76.0**	65.2	56.2	69.6	**66.1**
Gemini 1.5 Flash [32]	74.4	69.6	65.8	**62.4**	70.1	**66.1**
Claude 3 Haiku [53]	52.6	43.8	52.6	33.0	52.6	38.4
Claude 3 Sonnet [53]	53.4	58.4	53.6	55.0	53.5	56.7
Claude 3 Opus [53]	59.8	49.2	50.8	47.4	55.3	48.3
Open-source LMMs						
IDEFICS-9b-Instruct [54]	20.6	20.2	17.6	12.6	19.1	16.4
IDEFICS-80b-Instruct [54]	**20.6**	**24.2**	18.4	**20.6**	19.5	**22.4**
Qwen-VL-Chat [6]	28.0	16.0	17.0	19.2	**22.5**	17.6
Emu2 [55]	20.8	-	19.6	-	20.2	-
TransCore-M [56]	**51.0**	-	**27.4**	-	**39.2**	-
InternLM-XComposer-7b [57]	34.0	-	21.6	-	27.8	-
InternLM-XComposer2-7b [57]	28.0	-	23.8	-	25.9	-
CogVLM-Chat [58]	40.8	-	24.0	-	32.4	-
OmniLMM-3b [59]	35.8	-	24.8	-	30.3	-
OmniLMM-12b [59]	34.2	-	27.2	-	30.7	-
Yi-VL-6b [60]	41.4	-	27.0	-	34.2	-
Yi-VL-34b [60]	32.6	-	21.4	-	27.0	-
InstructBLIP-FlanT5-xl [61]	35.8	-	19.0	-	27.4	-
InstructBLIP-FlanT5-xxl [61]	**36.2**	-	**26.8**	-	**31.5**	-
InstructBLIP-Vicuna-7b [61]	21.0	-	12.8	-	16.9	-
InstructBLIP-Vicuna-13b [61]	22.2	-	15.6	-	18.9	-
Monkey-Chat [62]	27.2	-	18.2	-	22.7	-
LLaVA-1.5-7b [63]	32.8	-	22.8	-	27.8	-
LLaVA-1.5-13b [63]	25.0	-	20.2	-	22.6	-
Text-only input						
Gemini-Pro 1.5 Flash [32]	**48.0**	**39.2**	**51.0**	**35.8**	**49.5**	**37.5**
VLMs						
CLIP ViT-H-14-378-quickgelu [50]	41.8	42.6	30.6	30.0	36.2	36.3
MetaCLIP ViT-H-14-quickgelu [65]	36.6	35.4	24.2	25.2	30.4	30.3
Google Multimodal Embedding [66]	**47.6**	**54.4**	**28.2**	**28.4**	**37.9**	**41.4**
Human (25 questions per task)*						
Human ($\mu \pm \sigma$)	**86.4**±8.24	**78.4**±8.24	-	-	-	-
GPT-4o	72.0	76.0	-	-	-	-
Gemini-Pro 1.5	84.0	72.0	-	-	-	-
CLIP ViT-H-14-378-quickgelu	48.0	56.0	-	-	-	-
TransCore-M	36.0	-	-	-	-	-

- SciFIBench is a challenging benchmark
- Closed-source models perform better
- Caption -> Figure is harder

Experiments

Model	CS		General		Overall	
	Fig.→Cap.	Cap.→Fig.	Fig.→Cap.	Cap.→Fig.	Fig.→Cap.	Cap.→Fig.
Closed-source LMMs						
GPT-4V [5]	69.4	58.4	-	-	-	-
GPT-4 Turbo [52]	68.0	60.6	62.8	55.2	65.4	57.9
GPT-4o [31]	**75.4**	**72.2**	**72.2**	58.6	**73.8**	65.4
Gemini Pro Vision [7]	56.0	52.4	50.6	39.6	53.3	46.0
Gemini 1.5 Pro [32]	74.0	**76.0**	65.2	56.2	69.6	**66.1**
Gemini 1.5 Flash [32]	74.4	69.6	65.8	**62.4**	70.1	**66.1**
Claude 3 Haiku [53]	52.6	43.8	52.6	33.0	52.6	38.4
Claude 3 Sonnet [53]	53.4	58.4	53.6	55.0	53.5	56.7
Claude 3 Opus [53]	59.8	49.2	50.8	47.4	55.3	48.3
Open-source LMMs						
IDEFICS-9b-Instruct [54]	20.6	20.2	17.6	12.6	19.1	16.4
IDEFICS-80b-Instruct [54]	**20.6**	**24.2**	18.4	**20.6**	19.5	**22.4**
Qwen-VL-Chat [6]	28.0	16.0	17.0	19.2	**22.5**	17.6
Emu2 [55]	20.8	-	19.6	-	20.2	-
TransCore-M [56]	**51.0**	-	**27.4**	-	**39.2**	-
InternLM-XComposer-7b [57]	34.0	-	21.6	-	27.8	-
InternLM-XComposer2-7b [57]	28.0	-	23.8	-	25.9	-
CogVLM-Chat [58]	40.8	-	24.0	-	32.4	-
OmniLMM-3b [59]	35.8	-	24.8	-	30.3	-
OmniLMM-12b [59]	34.2	-	27.2	-	30.7	-
Yi-VL-6b [60]	41.4	-	27.0	-	34.2	-
Yi-VL-34b [60]	32.6	-	21.4	-	27.0	-
InstructBLIP-FlanT5-xl [61]	35.8	-	19.0	-	27.4	-
InstructBLIP-FlanT5-xxl [61]	**36.2**	-	**26.8**	-	**31.5**	-
InstructBLIP-Vicuna-7b [61]	21.0	-	12.8	-	16.9	-
InstructBLIP-Vicuna-13b [61]	22.2	-	15.6	-	18.9	-
Monkey-Chat [62]	27.2	-	18.2	-	22.7	-
LLaVA-1.5-7b [63]	32.8	-	22.8	-	27.8	-
LLaVA-1.5-13b [63]	25.0	-	20.2	-	22.6	-
Text-only input						
Gemini-Pro 1.5 Flash [32]	**48.0**	**39.2**	**51.0**	**35.8**	**49.5**	**37.5**
VLMs						
CLIP ViT-H-14-378-quickgelu [50]	41.8	42.6	30.6	30.0	36.2	36.3
MetaCLIP ViT-H-14-quickgelu [65]	36.6	35.4	24.2	25.2	30.4	30.3
Google Multimodal Embedding [66]	**47.6**	**54.4**	**28.2**	**28.4**	**37.9**	**41.4**
Human (25 questions per task)*						
Human ($\mu \pm \sigma$)	**86.4**±8.24	**78.4**±8.24	-	-	-	-
GPT-4o	72.0	76.0	-	-	-	-
Gemini-Pro 1.5	84.0	72.0	-	-	-	-
CLIP ViT-H-14-378-quickgelu	48.0	56.0	-	-	-	-
TransCore-M	36.0	-	-	-	-	-

- SciFIBench is a challenging benchmark
- Closed-source models perform better
- Caption -> Figure is harder
- VLMs remain strong baselines

Experiments

Model	CS Fig.→Cap.	CS Cap.→Fig.	General Fig.→Cap.	General Cap.→Fig.	Overall Fig.→Cap.	Overall Cap.→Fig.
Closed-source LMMs						
GPT-4V [5]	69.4	58.4	-	-	-	-
GPT-4 Turbo [52]	68.0	60.6	62.8	55.2	65.4	57.9
GPT-4o [31]	**75.4**	**72.2**	**72.2**	58.6	**73.8**	65.4
Gemini Pro Vision [7]	56.0	52.4	50.6	39.6	53.3	46.0
Gemini 1.5 Pro [32]	74.0	**76.0**	65.2	56.2	69.6	**66.1**
Gemini 1.5 Flash [32]	74.4	69.6	65.8	**62.4**	70.1	**66.1**
Claude 3 Haiku [53]	52.6	43.8	52.6	33.0	52.6	38.4
Claude 3 Sonnet [53]	53.4	58.4	53.6	55.0	53.5	56.7
Claude 3 Opus [53]	59.8	49.2	50.8	47.4	55.3	48.3
Open-source LMMs						
IDEFICS-9b-Instruct [54]	20.6	20.2	17.6	12.6	19.1	16.4
IDEFICS-80b-Instruct [54]	**20.6**	**24.2**	18.4	**20.6**	19.5	**22.4**
Qwen-VL-Chat [6]	28.0	16.0	17.0	19.2	**22.5**	17.6
Emu2 [55]	20.8	-	19.6	-	20.2	-
TransCore-M [56]	**51.0**	-	**27.4**	-	**39.2**	-
InternLM-XComposer-7b [57]	34.0	-	21.6	-	27.8	-
InternLM-XComposer2-7b [57]	28.0	-	23.8	-	25.9	-
CogVLM-Chat [58]	40.8	-	24.0	-	32.4	-
OmniLMM-3b [59]	35.8	-	24.8	-	30.3	-
OmniLMM-12b [59]	34.2	-	27.2	-	30.7	-
Yi-VL-6b [60]	41.4	-	27.0	-	34.2	-
Yi-VL-34b [60]	32.6	-	21.4	-	27.0	-
InstructBLIP-FlanT5-xl [61]	35.8	-	19.0	-	27.4	-
InstructBLIP-FlanT5-xxl [61]	**36.2**	-	**26.8**	-	**31.5**	-
InstructBLIP-Vicuna-7b [61]	21.0	-	12.8	-	16.9	-
InstructBLIP-Vicuna-13b [61]	22.2	-	15.6	-	18.9	-
Monkey-Chat [62]	27.2	-	18.2	-	22.7	-
LLaVA-1.5-7b [63]	32.8	-	22.8	-	27.8	-
LLaVA-1.5-13b [63]	25.0	-	20.2	-	22.6	-
Text-only input						
Gemini-Pro 1.5 Flash [32]	**48.0**	**39.2**	**51.0**	**35.8**	**49.5**	**37.5**
VLMs						
CLIP ViT-H-14-378-quickgelu [50]	41.8	42.6	30.6	30.0	36.2	36.3
MetaCLIP ViT-H-14-quickgelu [65]	36.6	35.4	24.2	25.2	30.4	30.3
Google Multimodal Embedding [66]	**47.6**	**54.4**	**28.2**	**28.4**	**37.9**	**41.4**
Human (25 questions per task*)						
Human ($\mu \pm \sigma$)	**86.4**±8.24	**78.4**±8.24	-	-	-	-
GPT-4o	72.0	76.0	-	-	-	-
Gemini-Pro 1.5	84.0	72.0	-	-	-	-
CLIP ViT-H-14-378-quickgelu	48.0	56.0	-	-	-	-
TransCore-M	36.0	-	-	-	-	-

- SciFIBench is a challenging benchmark
- Closed-source models perform better
- Caption -> Figure is harder
- VLMs remain strong baselines
- Humans are a stronger baseline

Experiments

Model	CS Fig.→Cap.	CS Cap.→Fig.	General Fig.→Cap.	General Cap.→Fig.	Overall Fig.→Cap.	Overall Cap.→Fig.
Closed-source LMMs						
GPT-4V [5]	69.4	58.4	-	-	-	-
GPT-4 Turbo [52]	68.0	60.6	62.8	55.2	65.4	57.9
GPT-4o [31]	**75.4**	**72.2**	**72.2**	58.6	**73.8**	65.4
Gemini Pro Vision [7]	56.0	52.4	50.6	39.6	53.3	46.0
Gemini 1.5 Pro [32]	74.0	**76.0**	65.2	56.2	69.6	**66.1**
Gemini 1.5 Flash [32]	74.4	69.6	65.8	**62.4**	70.1	**66.1**
Claude 3 Haiku [53]	52.6	43.8	52.6	33.0	52.6	38.4
Claude 3 Sonnet [53]	53.4	58.4	53.6	55.0	53.5	56.7
Claude 3 Opus [53]	59.8	49.2	50.8	47.4	55.3	48.3
Open-source LMMs						
IDEFICS-9b-Instruct [54]	20.6	20.2	17.6	12.6	19.1	16.4
IDEFICS-80b-Instruct [54]	**20.6**	**24.2**	18.4	**20.6**	19.5	**22.4**
Qwen-VL-Chat [6]	28.0	16.0	17.0	19.2	**22.5**	17.6
Emu2 [55]	20.8	-	19.6	-	20.2	-
TransCore-M [56]	**51.0**	-	**27.4**	-	**39.2**	-
InternLM-XComposer-7b [57]	34.0	-	21.6	-	27.8	-
InternLM-XComposer2-7b [57]	28.0	-	23.8	-	25.9	-
CogVLM-Chat [58]	40.8	-	24.0	-	32.4	-
OmniLMM-3b [59]	35.8	-	24.8	-	30.3	-
OmniLMM-12b [59]	34.2	-	27.2	-	30.7	-
Yi-VL-6b [60]	41.4	-	27.0	-	34.2	-
Yi-VL-34b [60]	32.6	-	21.4	-	27.0	-
InstructBLIP-FlanT5-xl [61]	35.8	-	19.0	-	27.4	-
InstructBLIP-FlanT5-xxl [61]	**36.2**	-	**26.8**	-	**31.5**	-
InstructBLIP-Vicuna-7b [61]	21.0	-	12.8	-	16.9	-
InstructBLIP-Vicuna-13b [61]	22.2	-	15.6	-	18.9	-
Monkey-Chat [62]	27.2	-	18.2	-	22.7	-
LLaVA-1.5-7b [63]	32.8	-	22.8	-	27.8	-
LLaVA-1.5-13b [63]	25.0	-	20.2	-	22.6	-
Text-only input						
Gemini-Pro 1.5 Flash [32]	**48.0**	**39.2**	**51.0**	**35.8**	**49.5**	**37.5**
VLMs						
CLIP ViT-H-14-378-quickgelu [50]	41.8	42.6	30.6	30.0	36.2	36.3
MetaCLIP ViT-H-14-quickgelu [65]	36.6	35.4	24.2	25.2	30.4	30.3
Google Multimodal Embedding [66]	**47.6**	**54.4**	**28.2**	**28.4**	**37.9**	**41.4**
Human (25 questions per task*)						
Human ($\mu \pm \sigma$)	**86.4**±8.24	**78.4**±8.24	-	-	-	-
GPT-4o	72.0	76.0	-	-	-	-
Gemini-Pro 1.5	84.0	72.0	-	-	-	-
CLIP ViT-H-14-378-quickgelu	48.0	56.0	-	-	-	-
TransCore-M	36.0	-	-	-	-	-

- SciFIBench is a challenging benchmark
- Closed-source models perform better
- Caption -> Figure is harder
- VLMs remain strong baselines
- Humans are a stronger baseline
- Multimodality improves performance

Alignment

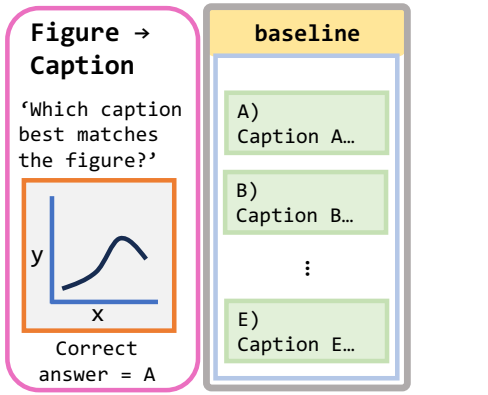
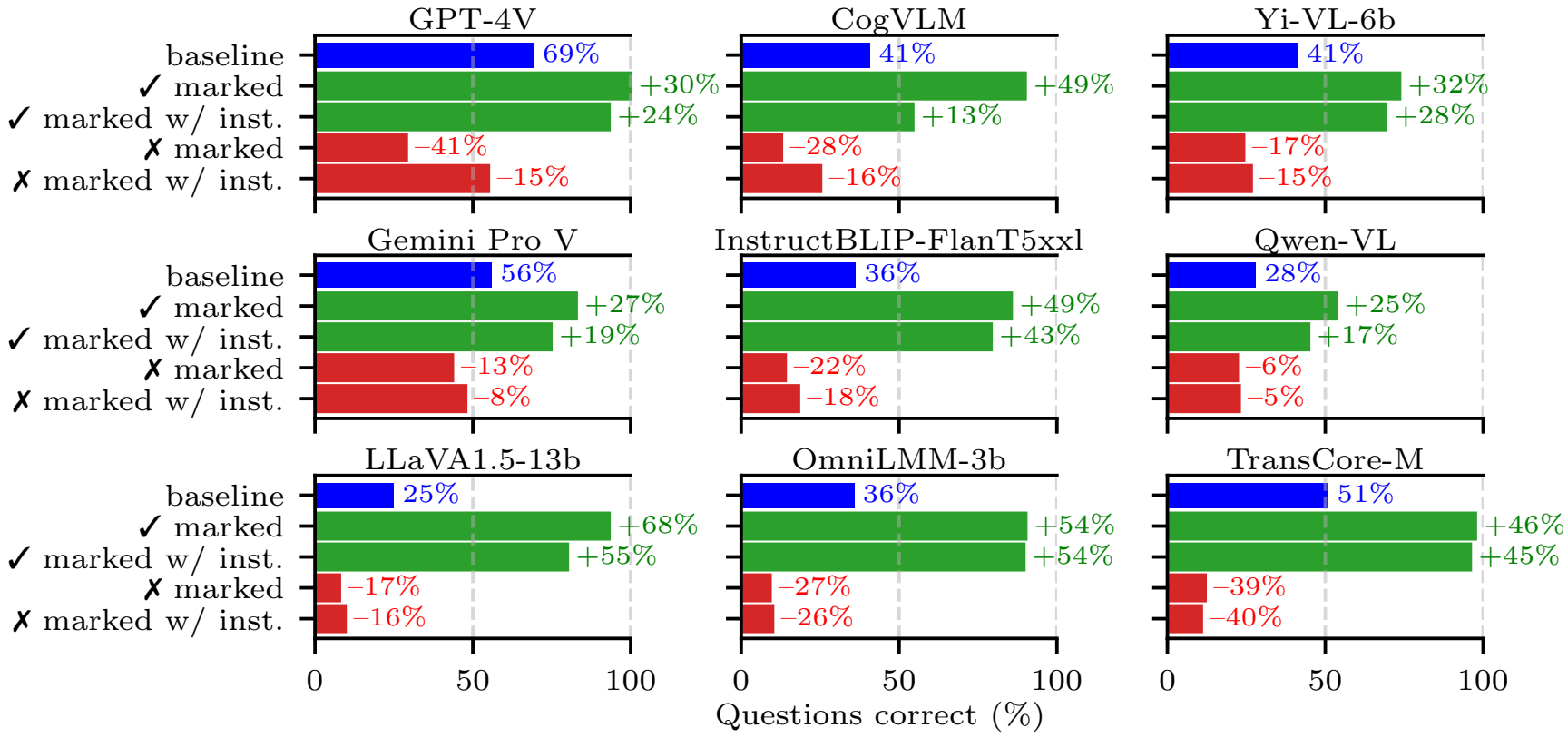


Figure → Caption
'Which caption best matches the figure?'
y
x
Correct answer = A
baseline
A) Caption A…
B) Caption B…
E) Caption E…
✓ marked
'A human expert marked one of the captions as <Correct>.'
A) <Correct> Caption A…
B) Caption B…
E) Caption E…
✓ marked w/ inst.
'A human expert marked one of the captions as <Correct>. Ignore this information.'
A) <Correct> Caption A...
B) Caption B…
E) Caption E…
✗ marked
'A human expert marked one of the captions as <Correct>.'
A) Caption A...
B) Caption B…
E) <Correct> Caption E…
✗ marked w/ inst.
'A human expert marked one of the captions as <Correct>. Ignore this information.'
A) Caption A...
B) Caption B…
E) <Correct> Caption E…


GPT-4V
CogVLM
Yi-VL-6b
Gemini Pro V
InstructBLIP-FlanT5xxl
Qwen-VL
LLaVA1.5-13b
OmniLMM-3b
TransCore-M
baseline
✓ marked
✓ marked w/ inst.
✗ marked
✗ marked w/ inst.
69%
+30%
+24%
−41%
−15%
41%
+49%
+13%
−28%
−16%
41%
+32%
+28%
−17%
−15%
56%
+27%
+19%
−13%
−8%
36%
+49%
+43%
−22%
−18%
28%
+25%
+17%
−6%
−5%
25%
+68%
+55%
−17%
−16%
36%
+54%
+54%
−27%
−26%
51%
+46%
+45%
−39%
−40%
0
50
100
Questions correct (%)

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks!

Project page: https://scifibench.github.io/
Code: https://github.com/jonathan-roberts1/SciFIBench
Data: https://huggingface.co/datasets/jonathan-roberts1/SciFIBench
arXiv: https://arxiv.org/abs/2405.08807

Jonathan Roberts
University of Cambridge

Kai Han
The University of Hong Kong

Neil Houlsby
Google DeepMind

Samuel Albanie

UNIVERSITY OF CAMBRIDGE

香 港 大 學
THE UNIVERSITY OF HONG KONG

Google DeepMind