Can Pre-trained Vision and Language Model Answer Visual Information-Seeking Questions?

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo,
Alan Ritter, Ming-Wei Chang

Georgia Institute of Technology
Google Research
Google DeepMind

InfoSeek: A new VQA Benchmark focuses on Visual Information-Seeking Questions to test World Knowledge in MLLMs.

- 1M + auto & 8.9k human annotated
- 10k + entities

Scan the QR code to access our data and paper

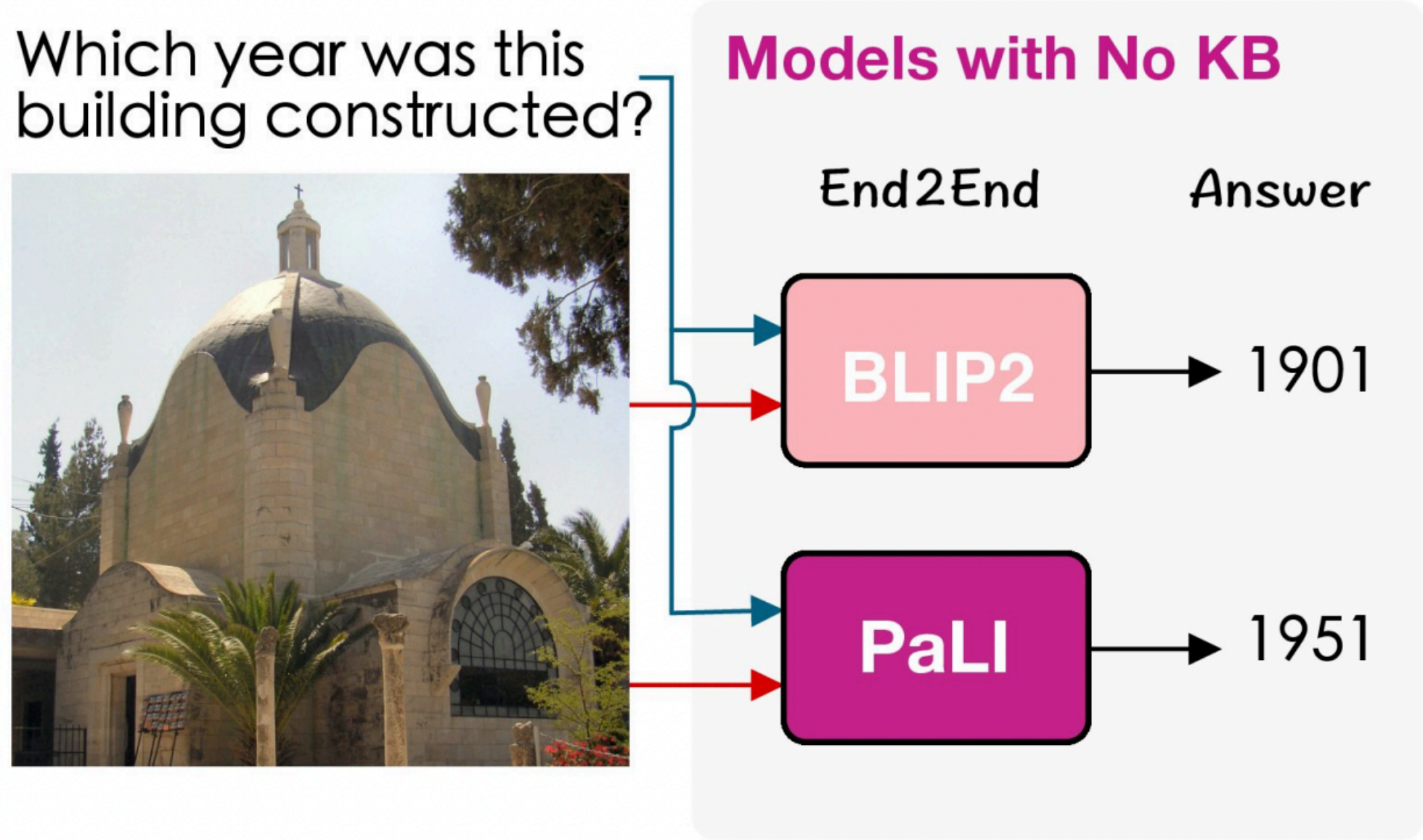
Q: What days might I most commonly go to this building?
A: Sunday
Previous VQA

Q: Who designed this building?
A: Antonio Barluzzi

Q: Which year was this building constructed?
A: 1955
INFOSEEK

- OK-VQA: 70.8% questions can be answered by average adults without using a search engine
- InfoSeek: only 4.4%. Query fine-grained information about the visual entity (e.g., Dominus Flevit Church)

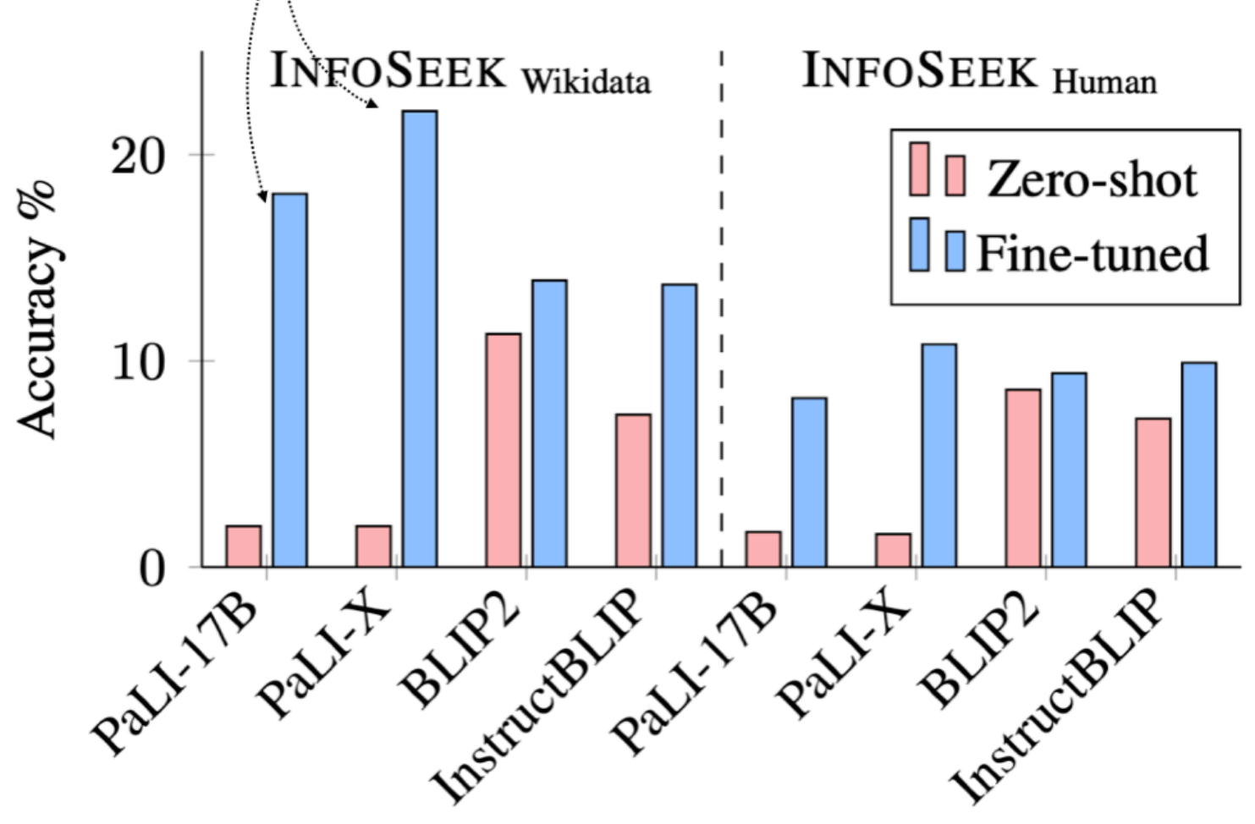
1. Goal: Benchmarking Visual World Knowledge in Multimodal LLMs



Model	LLM	# Params	INFOSEEK Overall	OK-VQA Accuracy	VQAv2 Accuracy
Random	-	-	0.1	-	-
Prior	-	-	3.2	-	32.1 †
PaLM (Q-only)	PaLM	540B	4.3	23.8	43.0
BLIP2	Flan-T5$_{XXL}$	12B	13.9	54.7	82.3
InstructBLIP	Flan-T5$_{XXL}$	12B	13.7	55.5	-
PaLI-17B	mT5$_{XXL}$	17B	18.1	64.8	84.6
PaLI-X	UL2$_{32B}$	55B	22.1	66.1	86.1

PaLI-X on InfoSeek vs other VQA datasets

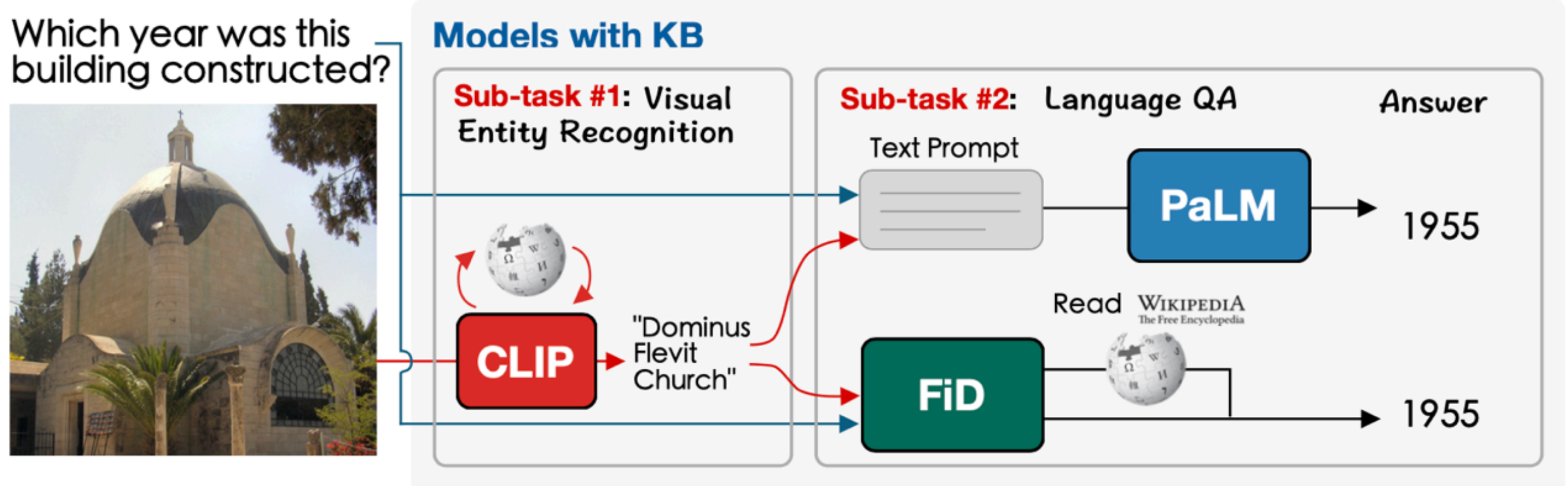
Eliciting knowledge from PaLI models by fine-tuning on InfoSeek



Instruction-tuned model makes less fine-grained predictions

Q: Which body of water is this mountain located in or next to?
A: Lake Como
`BLIP2`$_{(0\text{-shot})}$: lake como ✓
`InstructBLIP`$_{(0\text{-shot})}$: lake ✗

2. RAG: Retrieval-augmented models (CLIP + PaLM / Fusion-in Decoder)


Which year was this building constructed?
Models with KB
Sub-task #1: Visual Entity Recognition
Sub-task #2: Language QA
Answer
Text Prompt
PaLM
1955
Read WIKIPEDIA The Free Encyclopedia
CLIP
"Dominus Flevit Church"
FiD
1955

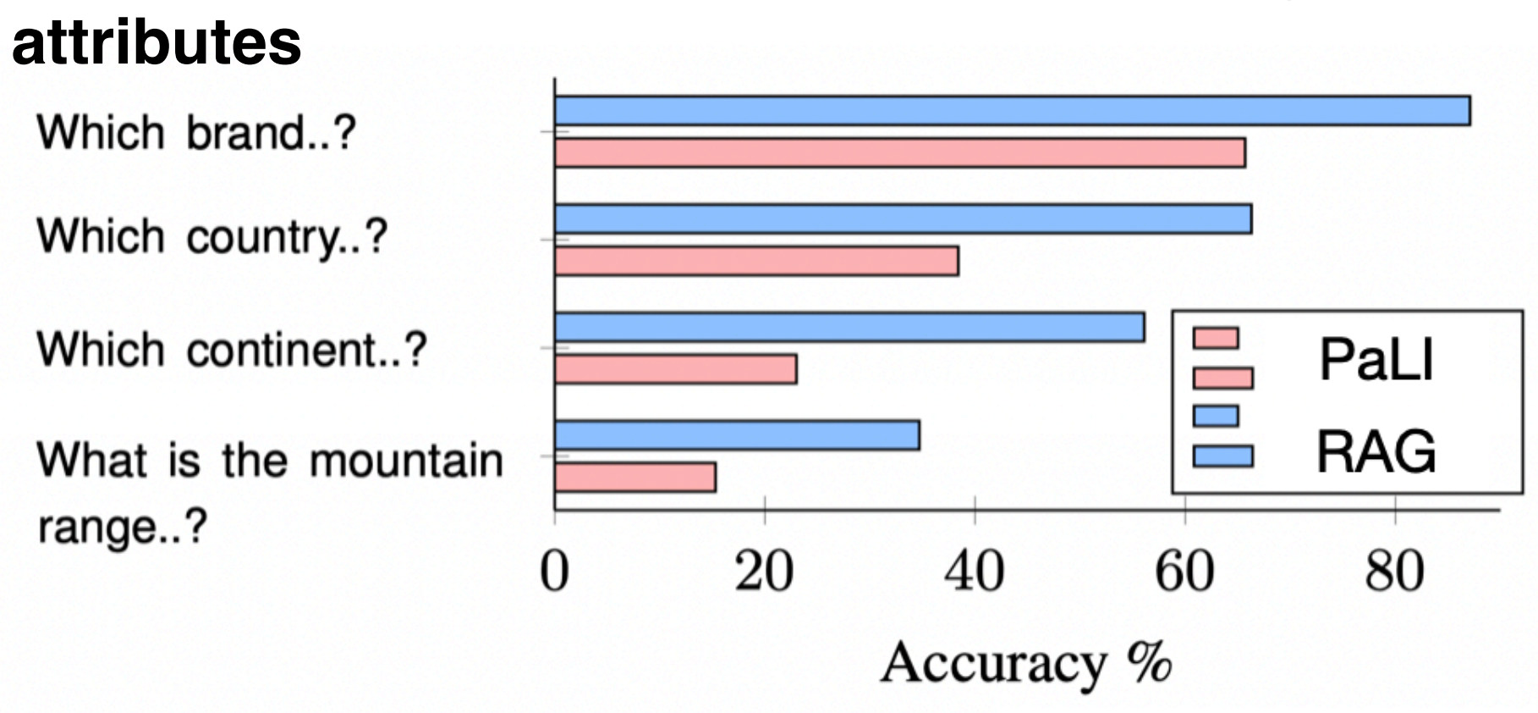
Model	TIME (Acc.)	NUMERICAL (Relaxed Acc.)	STRING (Acc.)
No-KB Setting			
Prior	0	4.4	5.0
PaLM (Q-only)	0	11.4	4.0
InstructBLIP	7.9	7.5	17.8
BLIP2	6.9	5.8	18.5
PaLI-17B	3.8	18.4	27.4
PaLI-X	7.7	16.1	30.0
With-KB Setting			
CLIP → PaLM	12.5	27.7	21.7
CLIP → FiD	12.3	23.4	23.9

Advantages on Time / Numerical questions

Model	INFOSEEK Wikidata	INFOSEEK Human	ENTITY Accuracy
Best No-KB	22.1	10.8	-
With-KB Setting			
CLIP → PaLM	20.1	15.2	22.2
CLIP → FID	19.3	18.2	
Oracle → FID	52.0	45.6	100

Big headroom for improvements

PaLI outperforms CLIP + FiD on less fine-grained attributes



Tail Entity → Head

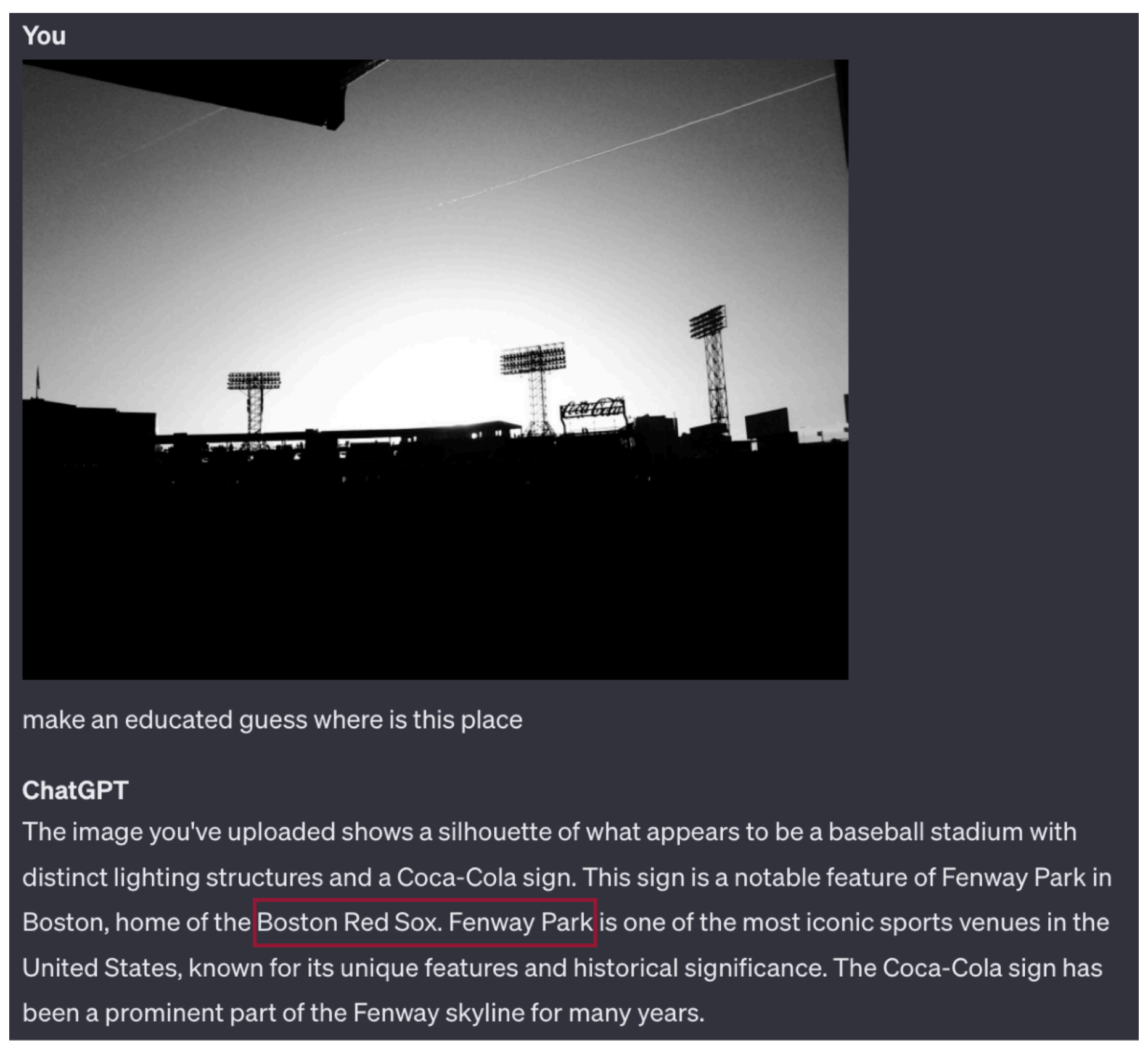
Q: Which region is this animal exclusively found?
A: U.S.A
End-to-end (PaLI): U.S.A ✓
Pipeline (Retrieval): Brazil ✗

Q: What country does this building belong to?
A: Brazil
End-to-end (PaLI): Brazil ✓
Pipeline (Retrieval): United Kingdom ✗

Q: What was the date this type of aircraft took the first flight?
A: 1981 September 26
End-to-end (PaLI): 1955 2 22 ✗
Pipeline (Retrieval): 1981 September 26 ✓

Q: What is the altitude of this building measured at meters above sea level?
A: 2430
End-to-end (PaLI): 3800 ✗
Pipeline (Retrieval): 2430 ✓

3.GPT-4 (Vision) on InfoSeek



You

make an educated guess where is this place

ChatGPT

The image you've uploaded shows a silhouette of what appears to be a baseball stadium with distinct lighting structures and a Coca-Cola sign. This sign is a notable feature of Fenway Park in Boston, home of the Boston Red Sox. Fenway Park is one of the most iconic sports venues in the United States, known for its unique features and historical significance. The Coca-Cola sign has been a prominent part of the Fenway skyline for many years.

Table 3: Benchmark results on VQA with fine-grained world knowledge. OC refers to Organization and Company. ♣ refers to that the corresponding model is trained using the training set of OK-VQA.

Model	Avg.	Building	Animal	Plant	Location	Food	OC	Facility	vehicle	Objects	Sport	Other
MiniGPT-4 (Vicuna-7b)	10.03	7.33	6.66	5.33	10.0	24.67	4.0	7.33	18.67	6.67	14.0	8.67
MiniGPT-4 (V-4-shot)	5.58	2.67	0.67	4.67	5.33	10.0	4.0	6.0	11.33	4.0	6.0	6.67
BLIP-2 (FlanT5-xxl)	10.67	8.7	2.67	4.0	16.0	14.0	9.33	16.0	28.0	2.0	9.33	7.33
BLIP-2 (Flan-4-shot)	9.44	5.3	2.0	13.33	14.0	8.0	12.0	10.67	28.67	6.67	13.33	6.67
InstructBLIP♣ (Vicuna-13b)	8.50	3.3	2.0	1.33	10.0	10.67	6.0	4.67	26.67	2.67	20.67	5.33
InstructBLIP♣ (V-4-shot)	8.75	5.3	2.0	2.0	8.0	15.33	4.67	5.33	16.67	5.33	24.67	6.67
InstructBLIP♣ (FlanT5-xxl)	8.37	4.0	5.33	2.0	8.67	8.0	8.0	8.0	28.0	5.34	8.67	6.0
InstructBLIP♣ (Flan-4-shot)	5.65	1.3	2.0	1.33	2.67	18.67	7.33	2.67	15.33	4.67	2.67	3.33
llava-v1.5-13b♣ (0-shot)	10.22	11.33	16.67	0.0	24.67	6.0	0.7	10.67	26.0	5.3	0.13	10.0
llava-v1.5-13b♣ (4-shot)	14.73	10.67	4.0	6.67	16.0	22.67	16.0	4.67	32.67	7.33	23.33	18.0
GPT-4V (0-shot)	**26.62**	18.67	10.28	12.60	17.44	46.67	19.33	29.33	33.33	17.69	57.33	22.53
GPT-4V (4-shot)	26.10	20.0	7.47	8.89	23.49	40.0	20.67	30.0	32.67	13.08	58.67	23.24

A Comprehensive Evaluation of GPT-4V on Knowledge-Intensive Visual Question Answering
Li et al., 2023, arXiv 2311.07536

4. Human Evaluation

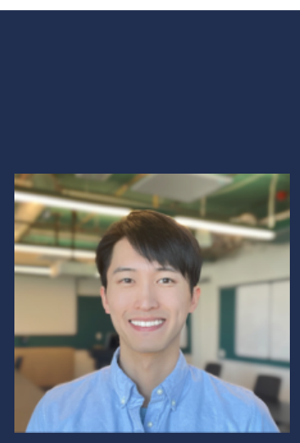
Which city is this hotel located in?
PALI: Pattaya ✓

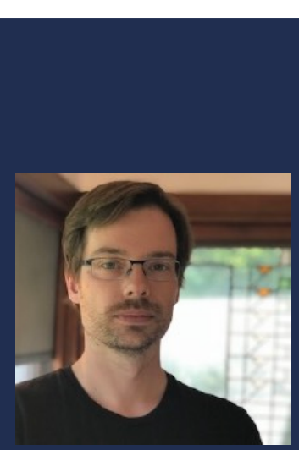
What year was this painting created?
PALI: 1884 ✓

What is the brand of this wheel?
PALI: schmiedmann ✗
Correct answer: HRE

Who is the designer of these shoes?
PALI: Nicolas Ghesquiere ✗
Correct answer: Virgil Abloh

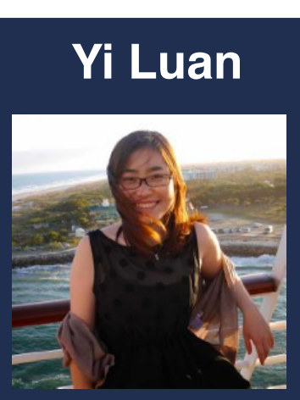
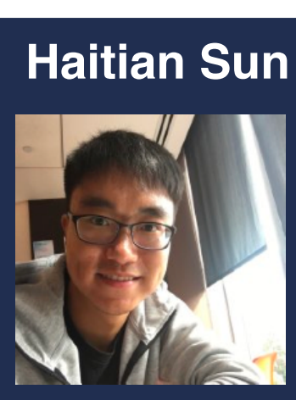
Georgia Tech

Yang Chen

yang.chen@cc.gatech.edu
@ychennip

Alan Ritter

alan.ritter@cc.gatech.edu
@alan_ritter

Google Research
Google DeepMind

Hexiang Hu

Yi Luan

Haitian Sun

Soravit Changpinyo

Ming-Wei Chang