

Visual Riddles:



a Commonsense and World Knowledge
Challenge for Large Vision and
Language Models



Nitzan Guetta



Aviv Slobodkin



Aviya Maimon



Eliya Habba



Royi Rassin



Yonatan Bitton



Amir Globerson



Idan Szpektor



Yuval Elovici

Why is he doing this?



What is this local doing?



Sara is a resort owner in Krabi, Thailand. could this be her resort?



Why is he doing this?



Look at the nightstand

What is this local doing?



Sara is a resort owner in Krabi, Thailand. could this be her resort?



Why is he doing this?

The image depicts a man scratching his arm, in a bedroom and a mosquito on a nightstand near the bed. Therefore, the man probably scratching his arm due to mosquito bite.

What is this local doing?



Sara is a resort owner in Krabi, Thailand. could this be her resort?



Why is he doing this?

The image depicts a man scratching his arm, in a bedroom and a mosquito on a nightstand near the bed. Therefore, the man probably scratching his arm due to mosquito bite.

What is this local doing?



Where is he?

Sara is a resort owner in Krabi, Thailand. could this be her resort?



Why is he doing this?

The image depicts a man scratching his arm, in a bedroom and a mosquito on a nightstand near the bed. Therefore, the man probably scratching his arm due to mosquito bite.

What is this local doing?

This local appears to be eating and pushing his finger to his cheek. In Italy, while eating, this gesture usually means “buono” - that you find the food tasty. Therefore, he is most likely saying that the food is delicious.

Sara is a resort owner in Krabi, Thailand. could this be her resort?



Why is he doing this?

The image depicts a man scratching his arm, in a bedroom and a mosquito on a nightstand near the bed. Therefore, the man probably scratching his arm due to mosquito bite.

What is this local doing?

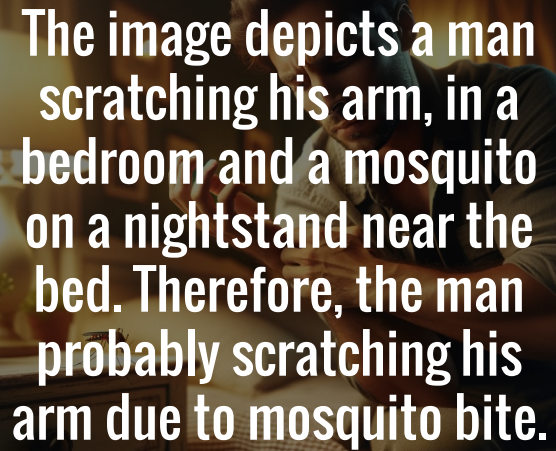
This local appears to be eating and pushing his finger to his cheek. In Italy, while eating, this gesture usually means “buono” - that you find the food tasty. Therefore, he is most likely saying that the food is delicious.

Sara is a resort owner in Krabi, Thailand. could this be her resort?



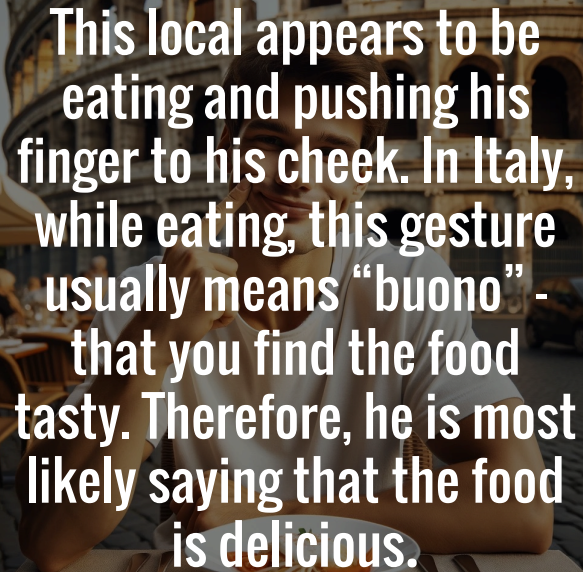
Look at the mountains

Why is he doing this?

A photograph of a man sitting on a bed in a bedroom, scratching his arm. A mosquito is visible on a nightstand near the bed.

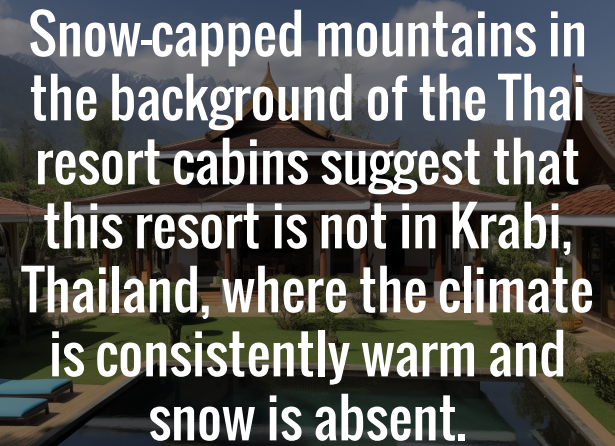
The image depicts a man scratching his arm, in a bedroom and a mosquito on a nightstand near the bed. Therefore, the man probably scratching his arm due to mosquito bite.

What is this local doing?

A photograph of a man sitting at a table eating, pushing his finger to his cheek.

This local appears to be eating and pushing his finger to his cheek. In Italy, while eating, this gesture usually means “buono” - that you find the food tasty. Therefore, he is most likely saying that the food is delicious.

Sara is a resort owner in Krabi, Thailand. could this be her resort?

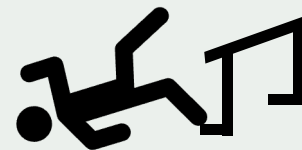
A photograph of a resort with snow-capped mountains in the background and Thai resort cabins.

Snow-capped mountains in the background of the Thai resort cabins suggest that this resort is not in Krabi, Thailand, where the climate is consistently warm and snow is absent.

Visual Riddles Challenge



Humans



V&L Models

Data Collection

Chocolate croissant
have a specific shape
in France



Data Collection

Chocolate croissant
have a specific shape
in France



What is the croissant
in the picture more
likely filled with -
chocolate or butter?

Data Collection

Chocolate croissant
have a specific shape
in France



Hint

Where is he?

What is the croissant
in the picture more
likely filled with -
chocolate or butter?

Data Collection

Chocolate croissant
have a specific shape
in France



What is the croissant
in the picture more
likely filled with -
chocolate or butter?

Hint

Where is he?

<https://en.wikipedia.org/wiki/Croissant>

Attribution

A screenshot of the Wikipedia article for "Croissant". The article text includes: "A croissant (UK: /kɹɔːsənt, kɹɔːsənt/; US: /kroːsɑːnt, kɹɔːsɑːnt/; French: [kʁwa.sɑ̃] ⓘ ⓘ ⓘ) is a French pastry made from puff pastry in a crescent shape.[2] It is a buttery, flaky, viennoiserie pastry inspired by the shape of the Austrian kipferl, but using the French yeast-leavened laminated dough.[3] Croissants are named for their historical crescent shape. The dough is layered with butter, rolled and folded several times in succession, then rolled into a thin sheet, in a technique called laminating. The process results in a layered, flaky texture, similar to a puff pastry. Crescent-shaped breads have been made since the Renaissance, and crescent-shaped cakes possibly since antiquity.[4] The modern croissant was developed in the early 20th century, when French bakers replaced the brioche dough of the kipferl with a yeast-leavened laminated dough.[5] In the late 1970s, the development of factory-made, frozen, preformed but unbaked dough made them into a fast food that could be freshly baked by unskilled labor. The croissant bakery, notably the La Croissanterie chain, was a French response to American-style fast food,[6] and as of 2008, 30–40% of the croissants sold in French bakeries and patisseries were baked from frozen dough.[7] Croissants are a common part of a continental breakfast in many European countries." The right sidebar shows a table with the following information: Type: Viennoiserie, Course: Breakfast, Place of origin: France, Main ingredients: Yeast-leavened dough, butter, Variations: Pain aux raisins, pain au chocolat, pain aux fraises. There is also a "Media: Croissant" link at the bottom.

Data Collection

Chocolate croissant
have a specific shape
in France



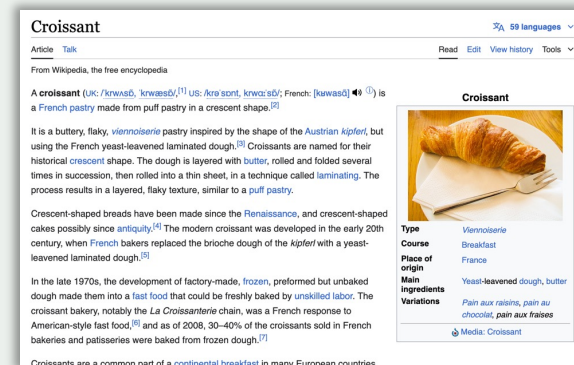
What is the croissant
in the picture more
likely filled with -
chocolate or butter?

Hint

Where is he?

<https://en.wikipedia.org/wiki/Croissant>

Attribution



The screenshot shows the Wikipedia article for 'Croissant'. The title is 'Croissant' and it is categorized as a 'Vienniserie'. The article text describes it as a buttery, flaky, yeast-leavened laminated dough, inspired by the shape of the Austrian kipferl. It mentions that croissants are named for their historical crescent shape and are typically filled with butter. The article also notes that croissants are a common part of a continental breakfast in many European countries.

Type	Vienniserie
Course	Breakfast
Place of origin	France
Main ingredients	Yeast-leavened dough, butter
Variations	Pain aux raisins, pain au chocolat, pain aux fraises

Answer

Given the Eiffel Tower in the background, we can assume this is Paris, France. In France, croissants are typically filled with butter (and also almonds), whereas the chocolate equivalent is called "Pain au Chocolat" and is more in a rectangular shape. Therefore, this croissant is most likely filled with butter, rather than chocolate.



Open-Ended VQA

+ Auxiliary Data

Q: What is probably the gender of the cat?

H: Look at the colors of the fur

At: https://en.wikipedia.org/wiki/Calico_cat

Calico cat

Article Talk

A calico cat (US English) is a domestic cat of any with large orange and black patches; however, they may have other colors in their patches. Calicoes are almost exclusively female, except under rare genetic conditions.

Multiple Choice

Choose the right answer out of five options:

1 GT answer, 3 Incorrect answers, 1 "cannot determine" distractor

Automatic Evaluation

Reference Free:

Given Q, is the candidate answer correct?

Reference Based:

Given Q and GT answer, is the candidate answer correct?



Open-Ended VQA

+ Auxiliary Data

Q: What is probably the gender of the cat?

H: Look at the colors of the fur

At: https://en.wikipedia.org/wiki/Calico_cat

Calico cat

Article Talk

A calico cat (US English) is a domestic cat of any with large orange and black patches; however, they may have other colors in their patches. Calicoes are almost exclusively female, except under rare genetic conditions.

Multiple Choice

Choose the right answer out of five options:

1 GT answer, 3 Incorrect answers, 1 "cannot determine" distractor

Automatic Evaluation

Reference Free:

Given Q, is the candidate answer correct?

Reference Based:

Given Q and GT answer, is the candidate answer correct?



Open-Ended VQA

+ Auxiliary Data

Q: What is probably the gender of the cat?

H: Look at the colors of the fur

At: https://en.wikipedia.org/wiki/Calico_cat



Calico cat

Article Talk

A calico cat (US English) is a domestic cat of any with large orange and black patches; however, they may have other colors in their patches. Calicoes are almost exclusively female, except under rare genetic conditions.

Multiple Choice

Choose the right answer out of five options:

1 GT answer, 3 Incorrect answers, 1 "cannot determine" distractor

Automatic Evaluation

Reference Free:

Given Q, is the candidate answer correct?

Reference Based:

Given Q and GT answer, is the candidate answer correct?



Open-Ended VQA

+ Auxiliary Data

Q: What is probably the gender of the cat?

H: Look at the colors of the fur

At: https://en.wikipedia.org/wiki/Calico_cat

Calico cat

Article Talk

A calico cat (US English) is a domestic cat of any with large orange and black patches; however, they may have other colors in their patches. Calicoes are almost exclusively female, except under rare genetic conditions.

Multiple Choice

Choose the right answer out of five options:

1 GT answer, 3 Incorrect answers, 1 "cannot determine" distractor

Automatic Evaluation

Reference Free:

Given Q, is the candidate answer correct?

Reference Based:

Given Q and GT answer, is the candidate answer correct?

Experiments: Open-ended VQA

		% Human Rating ↑
LVLM	Gemini Pro 1.5	40
	Gemini Pro Vision	30
	GPT4	34
	LLaVA-1.6-34B	15
	LLaVA-1.5-7B	13
	InstructBlip	13
Caption → LLM	Gemini Pro 1.5 → Gemini Pro 1.5	23
	Human (Oracle) → Gemini Pro 1.5	50
	Humans	82

Even with auxiliary data-model performance improved only marginally.

> Tests models' ability to generate correct answers from visual clues alone.

Experiments: Multiple-choice VQA

	% Accuracy ↑	% Cannot Determine	% Accuracy w/o Cannot Determine	+ Hint ↑	+ Attribution ↑
Gemini Pro 1.5	38	20	48	66	72
Gemini-Pro-Vision	41	3	42	62	-
GPT4	45	12	52	69	82
LLaVA-1.6-34B	24	8	26	30	-
LLaVA-1.5-7B	17	0	17	29	-

Importance
of auxiliary
information

> Evaluated automatically,
simple metric.

> Selecting the correct answer
out of five options.

Experiments:

Automatic Evaluation

	Judge	Accuracy of Judge Prediction Compared to Human Rating ↑
Reference-Based	Gemini Pro 1.5	87
	Gemini Pro Vision	75
	GPT4	86
	LLaVA -1.6-34b	76
	LLaVA -1.5-7b	70
Reference-Free	Gemini Pro 1.5	52
	Gemini Pro Vision	38
	GPT4	51
	LLaVA -1.6-34b	43
	LLaVA -1.5-7b	35

> For Open-Ended
VQA: Gemini Pro
1.5 + Ref-Based

	Visual Riddles		Hints	Attribution
	% Human Rating ↑	% Auto-Rater Rating ↑	% Auto-Rater Rating ↑	% Auto-Rater Rating ↑
Gemini Pro 1.5	40	53	62	29
Gemini Pro Vision	30	34	47	-
GPT4	34	38	61	25
LLaVA -1.6-34b	15	21	16	-
LLaVA -1.5-7b	13	19	30	-
InstructBlip	13	20	28	-
Humans	82	78	-	-

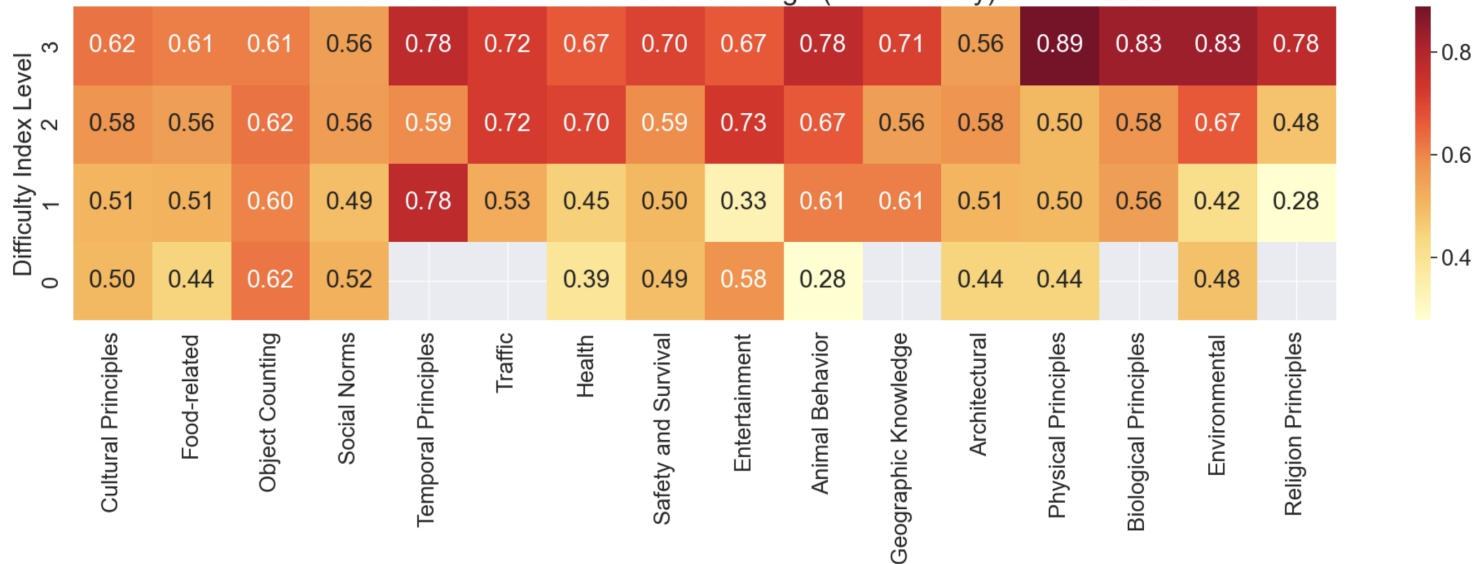
Ongoing
challenges
in model
reasoning with
auxiliary data

Commonsense Categorization of Visual Riddles

- Diversity and complexity of the reasoning skills.

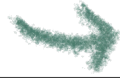
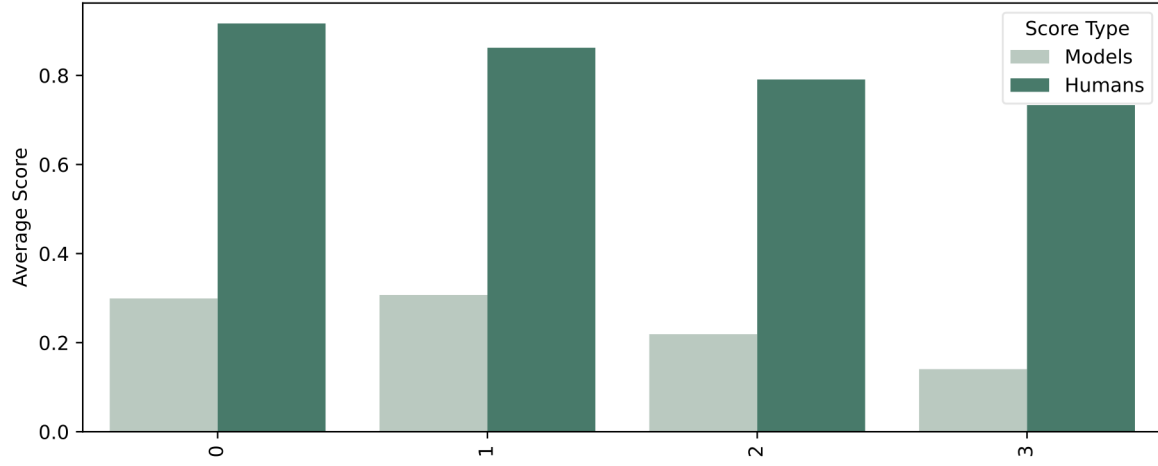
Animal Behavior	Architectural	Safety and Survival	Social Norms	Environmental
				
Who is probably a pet owner?	What will the man in the picture do if he wants to breathe fresh air?	Which chair should I sit on?	Is the man in the picture sitting alone?	He wants to go outside, what should he wear?

Models Fail to Solve Visual Riddles - Average (1 - Accuracy) of All Models

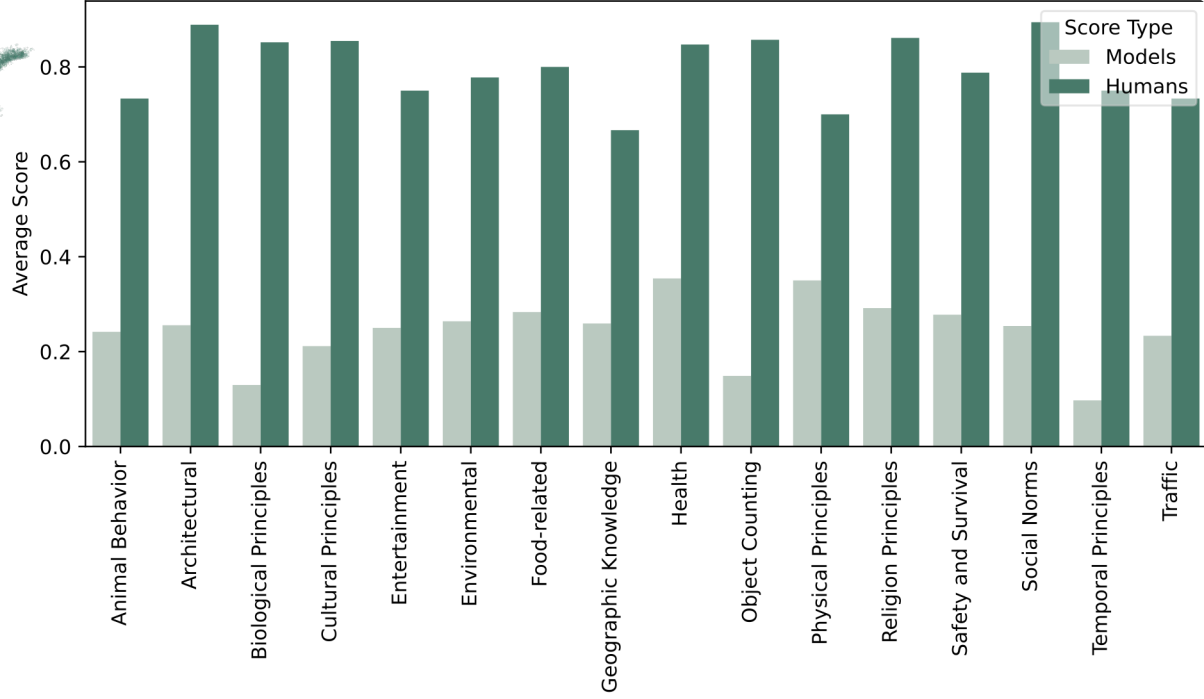


Models vs. Human Performance

Categories

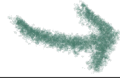
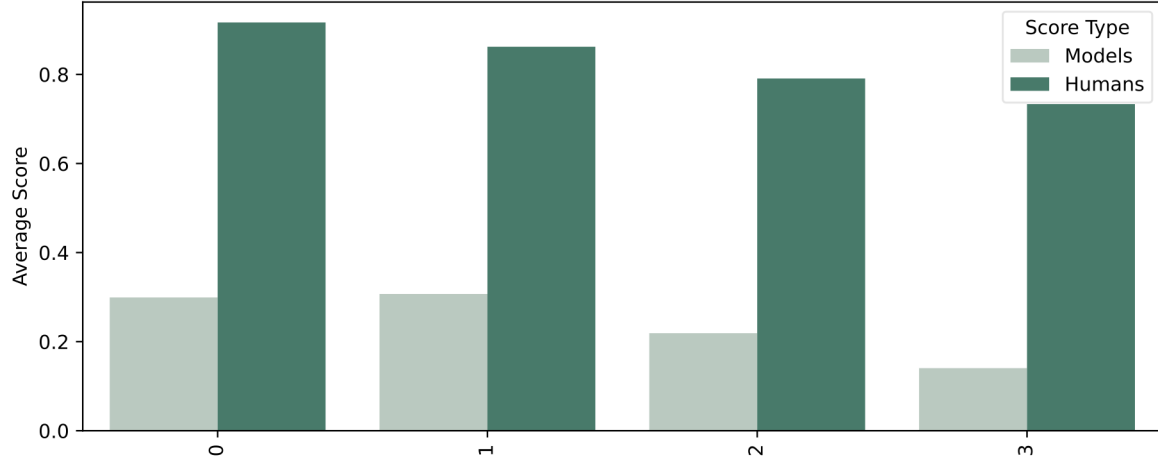


Difficulty
Levels

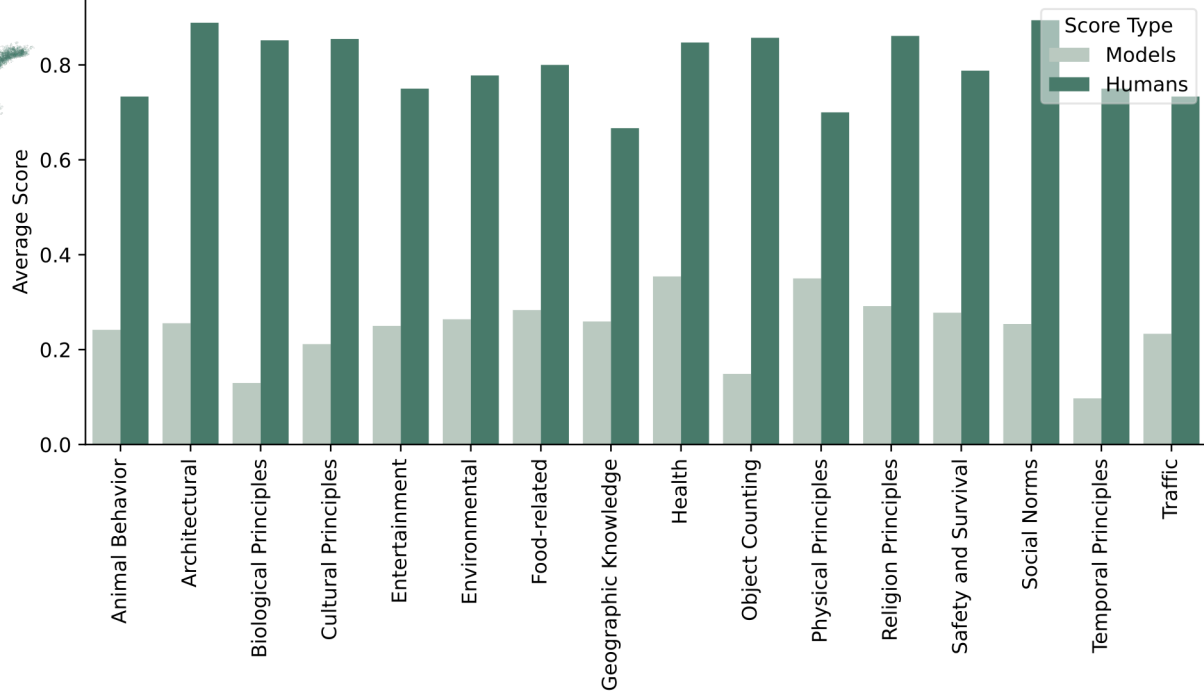
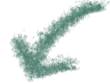


Models vs. Human Performance

Categories

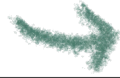
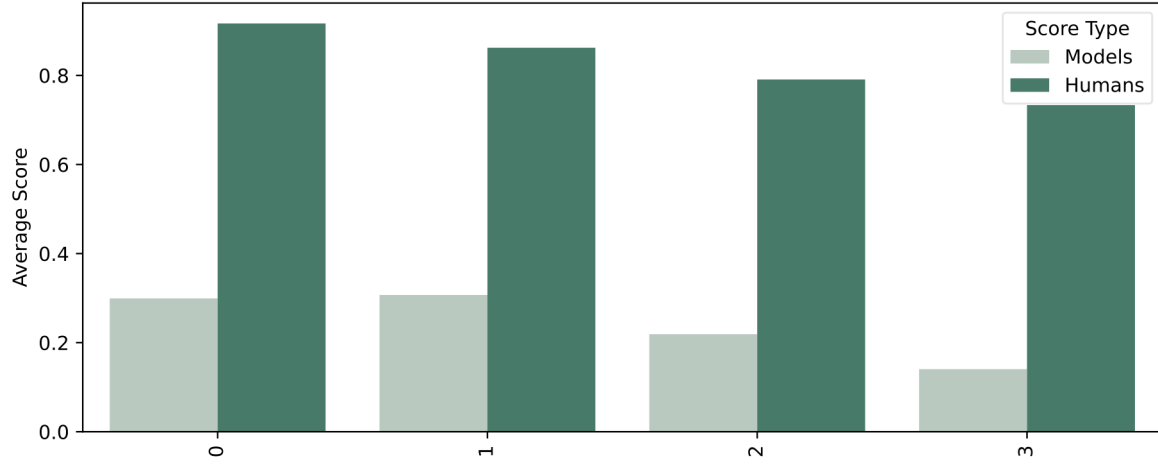


Difficulty
Levels

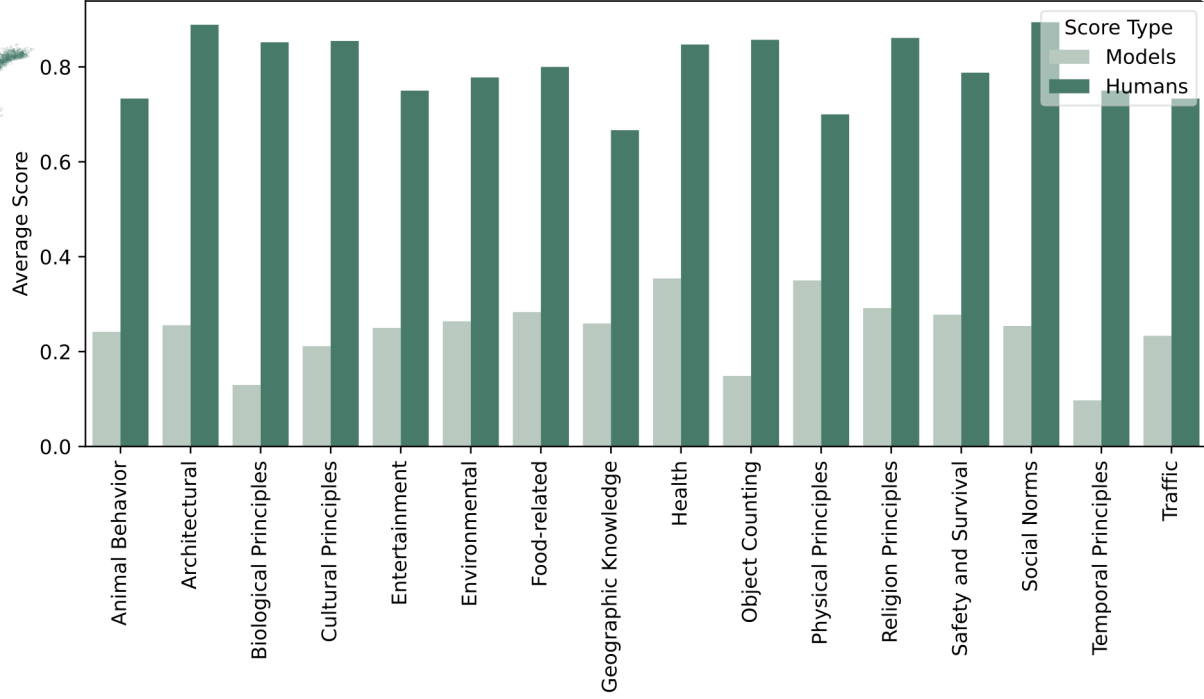


Models vs. Human Performance

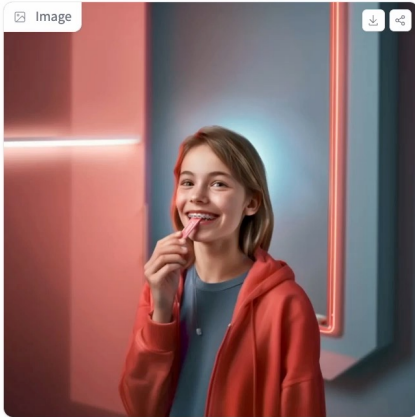
Categories



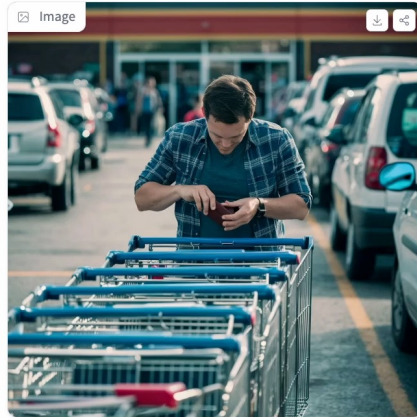
Difficulty
Levels



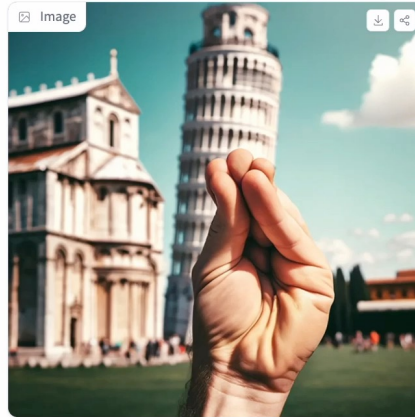
Slide to iterate Visual Riddles



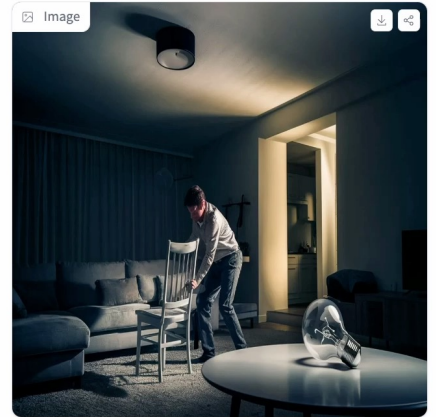
Click for details



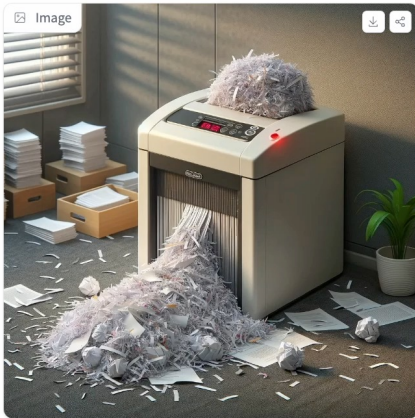
Click for details



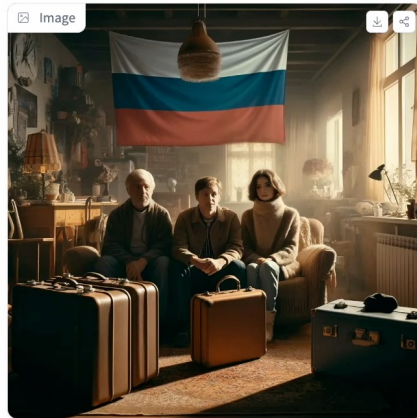
Click for details



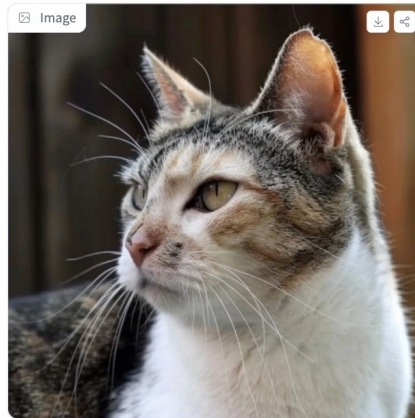
Click for details



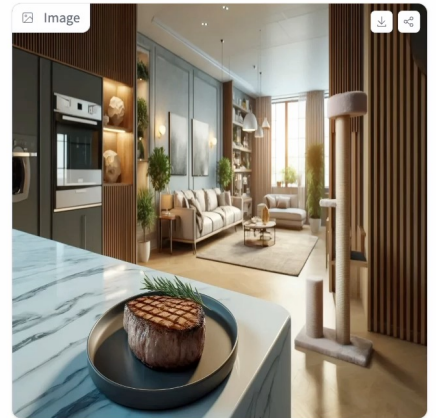
Click for details



Click for details



Click for details



Click for details



VISIT US ON VISUAL RIDDLES  WEBSITE:

<https://visual-riddles.github.io/>

