Models That Explain Themselves

Project 4: Multimodal Task with Prompting LLMs

Radhika Yadav
Valentina Tretti

Task + Dataset + Model

- **Task:** Visual Question Answering

- **Dataset:** A-OKVQA
  - Text: Validation split (Direct answers)
  - Images: COCO (Including captions)

- **Model:** “google/flan-t5-large”

Prompting: Answer prediction

```
Please answer the question given an image context. Use these examples for help.

Example 1
Image context: A bunch of bananas sitting on top of a wooden table.
Question: What treat can be made with this fruit and ice cream?
Answer: banana split

Example 2
Image context: A pan filled with onions sitting next to a pan of stew.
Question: The reddish-brown food in the further bowl is what type of food?
Answer: meat

Image context: A blue and white train is moving on the rails.
Question: What is the primary purpose for this train?
Answer:
```

Prompting: Explanation prediction

```
Please generate explanations for the given answer.

Image context: A blue and white train is moving on the rails.
Question: What is the primary purpose for this train?
Answer: to transport people
Explanation:
```

Scores

	METEOR
Answers	31.42%
Explanations	34.42%

Annotation scheme

- The ***System Causability Scale (SCS)*** proposed by Holzinger et al. (2020) measures ***how useful explanations are.***

- ***Explanation Goodness Checklist*** introduced by Hoffman et al. (2018) is used to look at a given explanation and make an a priori (or decontextualized) judgment as to ***whether or not it is "good."***

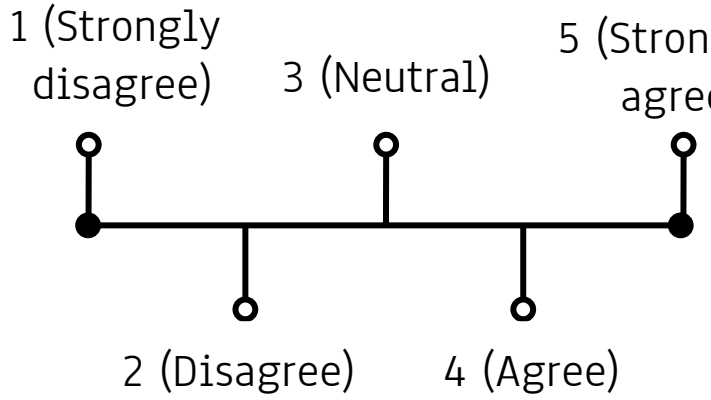
- The ***Explanation Satisfaction scheme*** by Hoffman et al. (2018) measures the ***degree to which users*** feel that they ***understand the explanation.*** It is a contextualized, a posteriori judgement of explanations.

Annotation scheme

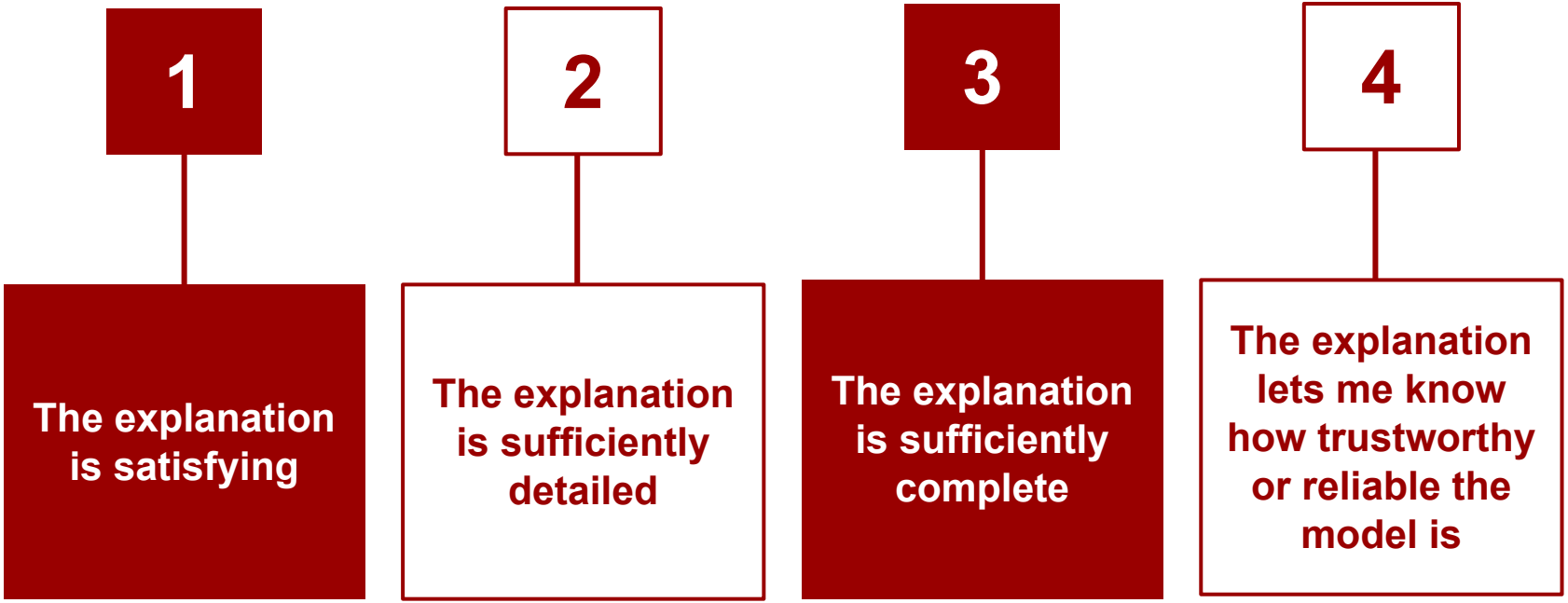
Goodness

0 (No) — 1 (Yes)

Satisfaction

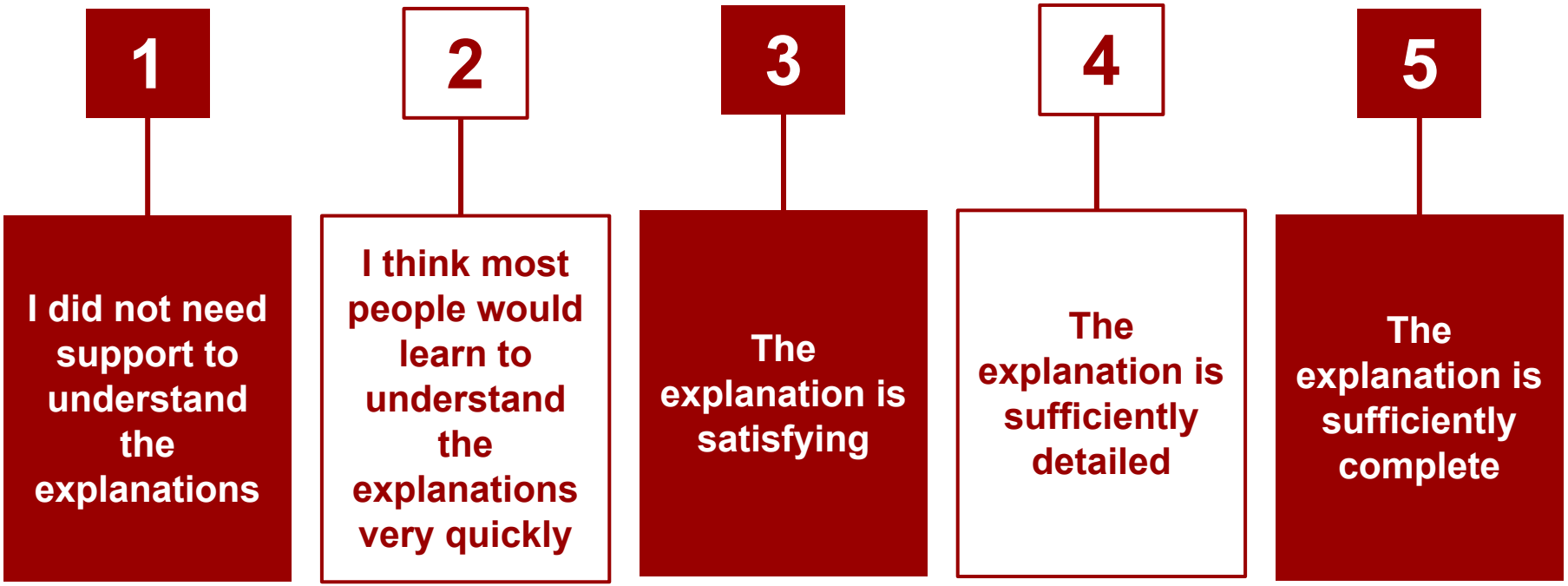


Annotations: Goodness



Adapted Explanation Goodness checklist by Hoffman et al. (2018)

Annotations: Satisfaction



System Causability Scale by Holzinger et al. (2020) + Explanation Satisfaction scheme by Hoffman et al. (2018)

Annotation scores and Agreement

	G1	G2	G3	G4	S1	S2	S3	S4	S5
Radhika	0.44	0.43	0.43	0.44	3.54	3.52	2.92	2.91	2.84
Valentina	0.44	0.53	0.82	0.52	3.34	3.73	2.82	2.95	3.72
Average	0.44	0.48	0.62	0.48	0.53	3.62	2.87	2.93	3.28

	G1	S3
IAA Score	0.63	0.27

Good predictions: Example 1

Image context

A crowd of people standing on snow covered ground.

Question

How are the skiers getting up the hill?

Answer

a ski lift

Explanation

The ski lift is used to get people up the hill.

Good predictions: Example 2

Image context

A man standing near the water with an umbrella.

Question

Why is the man holding an umbrella?

Answer

He is standing near the water

Explanation

The umbrella is used to protect the person standing near the water from the rain.

Bad predictions: Example 1

Image context

there is a bike that is parked on a small bridge

Question

Which mode of transport could traverse toward the horizon?

Answer

bike

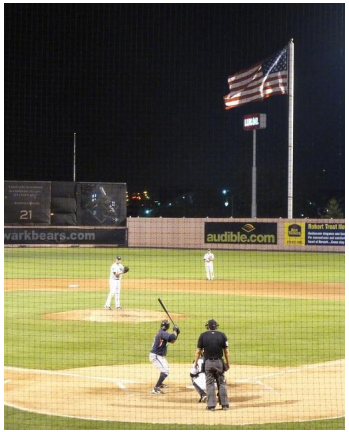
Explanation

The horizon is the line between the top of the mountain and the bottom of the valley

Bad predictions: Example 2

Image context

A small lit baseball field at night with people playing baseball.



Question

What country is there?

Answer

United States

Explanation

The United States is the only country that has a baseball field.

Interesting cases: Example 1

Image context

An elderly woman sitting on the bench resting.

Question

What is the age of this woman?

Answer

a woman who is elderly

Explanation

The woman is sitting on a bench resting. The age of the woman is unknown

Interesting cases: Example 2

Image context

Boats docked on land sitting side by side next to a lake.

Question

Why do these boats are painted in different color?

Answer

They are for different purposes

Explanation

The boats are painted in different colors because they are for different purposes.

Interesting cases: Example 2

Image context

Two men in bow ties standing next to steel rafter.

Question

What is the most likely explanation for how the people in the back of the room can hear this man?

Answer

They are in a room with acoustics.

Explanation

Acoustics is the property of a room that allows sound to travel through it.

Limitations and Learned Outcomes

- Captions don’t include enough information about the picture
- Lack of a standard annotation scheme

References

1. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., & Mottaghi, R. (2022, October). A-okvqa: A benchmark for visual question answering using world knowledge. In European Conference on Computer Vision (pp. 146-162). Cham: Springer Nature Switzerland.
2. Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations. *KI-Künstliche Intelligenz*, *34*(2), 193-198.
3. Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.
4. Hakimov, S., & Schlangen, D. (2023). Images in Language Space: Exploring the Suitability of Large Language Models for Vision & Language Tasks. arXiv preprint arXiv:2305.13782.