

Long-Form Answers to Visual Questions from Blind and Low Vision People

Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu,
Danna Gurari, Eunsol Choi, Amy Pavel

COLM 2024



Carnegie
Mellon
University





**2.2 Billion People with
Visual Impairments
Worldwide**

Short Answers to Visual Questions

Q. What is this?



Short Answers to Visual Questions

Q. *What is this?*



Short Answer

A. *Soy milk*

Short Answers to Visual Questions

Q. *What is this?*



Short Answer

A. *Soy milk*

What brand is this soy milk?

What flavor?

What is the nutritional information?



Long Answers to Visual Questions

Q. *What is this?*



Short Answer

A. *Soy milk*

Long Answer - GPT-4V

A. *The image you've uploaded appears to be of a carton of "Silk" brand chocolate soy milk. The packaging highlights that it's an all-natural soy milk product, with 50% more calcium than dairy milk, contains added vitamins, and has 150 calories per container. The size of the container is 8 fluid ounces (240 mL).*

Long Answers to Visual Questions

Q. *Can I wear these together?*



Long Answers to Visual Questions

Q. *Can I wear these together?* **Short Answer**

A. Yes



Long Answers to Visual Questions

Q. *Can I wear these together?*



Short Answer

A. Yes

Long Answer - GPT-4V

A. *Yes, you can wear these two pieces together. The image shows a thin strap top with cheetah print on it in golden, brown, and black colors. The outer cardigan is beige in color and seems to be made of wool. From a color perspective, there is harmony between the two.*

Long Answers to Visual Questions

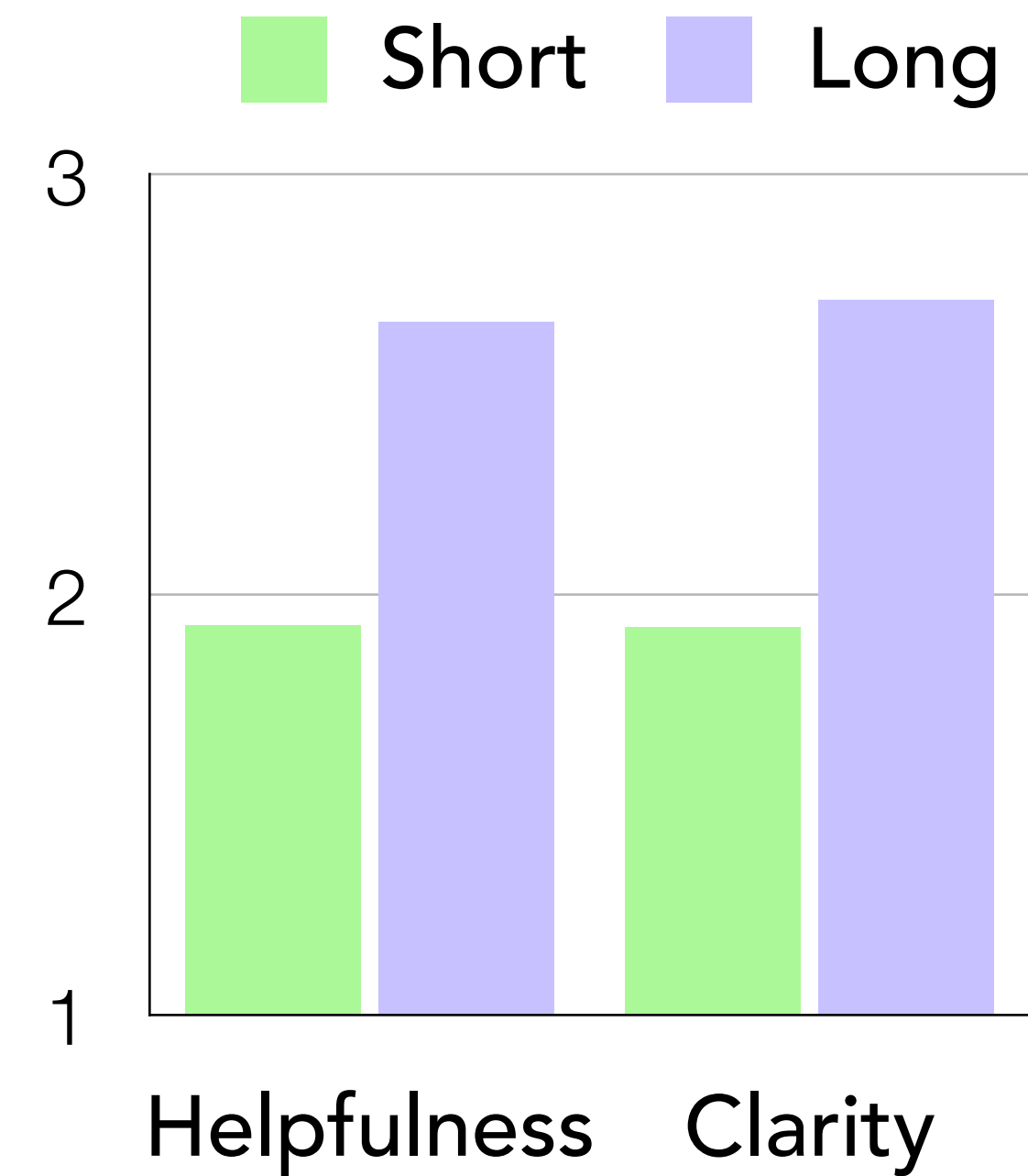


GPT-4V Powered Visual Assistant

Long Answers to Visual Questions



GPT-4V Powered Visual Assistant



Ratings from Blind People (N=8)

Understanding Long-Form Answers to Visual Questions from Blind and Low Vision People

Understanding Long-Form Answers to Visual Questions from Blind and Low Vision People

RQ1. *What information is provided in long-form answers?*

RQ2. *How can we evaluate long-form answers?*

RQ3. *How can we mitigate hallucinations in long-form answers?*

Understanding Long-Form Answers to Visual Questions from Blind and Low Vision People

RQ1. *What information is provided in long-form answers?*

RQ2. *How can we evaluate long-form answers?*

RQ3. *How can we mitigate hallucinations in long-form answers?*

Understanding Long-Form Answers to Visual Questions from Blind and Low Vision People

RQ1. *What information is provided in long-form answers?*

RQ2. *How can we evaluate long-form answers?*

RQ3. *How can we mitigate hallucinations in long-form answers?*

Understanding Long-Form Answers to Visual Questions from Blind and Low Vision People

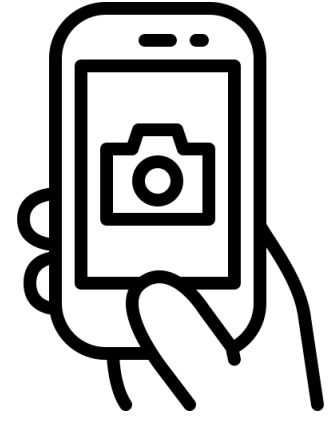
RQ1. *What information is provided in long-form answers?*

RQ2. *How can we evaluate long-form answers?*

RQ3. *How can we mitigate hallucinations in long-form answers?*

Dataset Collection

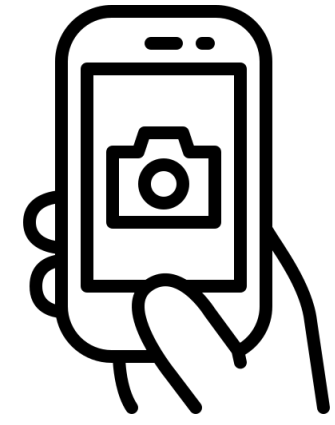
VizWiz (VizWiz, Gurari et al.)



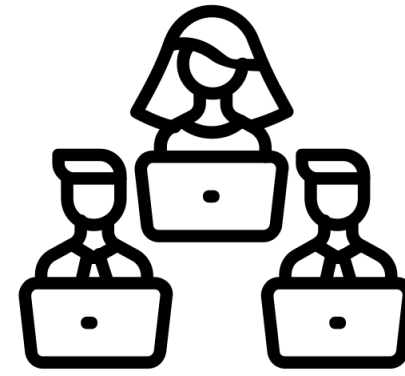
33k Image + Question Pairs
from blind people

Dataset Collection

VizWiz (VizWiz, Gurari et al.)



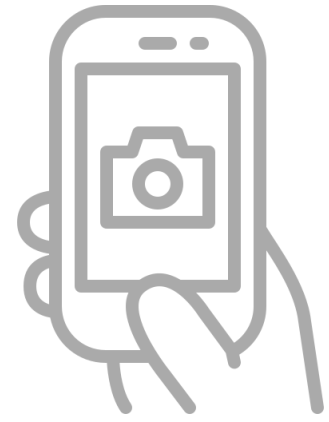
33k Image + Question Pairs
from blind people



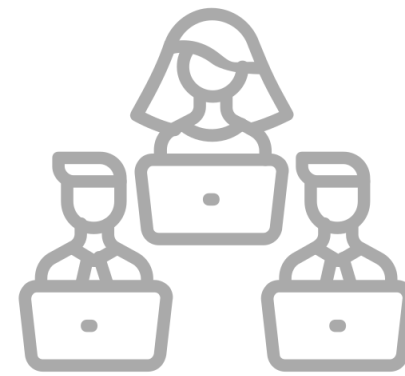
10 Short Answer per Question
from crowd workers

Dataset Collection

VizWiz (VizWiz, Gurari et al.)

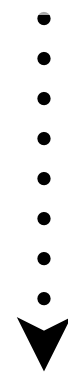


33k Image + Question Pairs
from blind people



10 Short Answer per Question
from crowd workers

VizWiz-LF (ours)



Sampled 600 Visual Questions

Identification

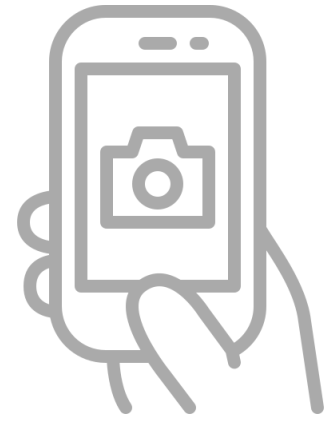
Description

Reading

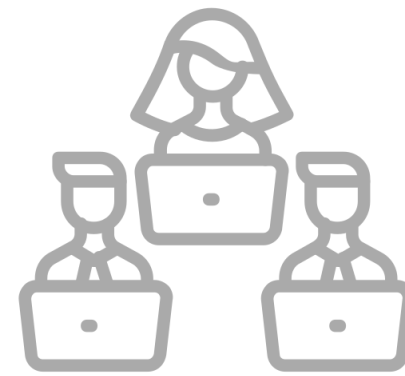
Misc.

Dataset Collection

VizWiz (VizWiz, Gurari et al.)



33k Image + Question Pairs
from blind people



10 Short Answer per Question
from crowd workers

VizWiz-LF (ours)

Sampled 600 Visual Questions>

4.2k Long-Form Answers
from experts and models

Identification

Description

Reading

Misc.

Human Experts

GPT-4V

InstructBLIP

Gemini

LLaVA

BLIP-2

QWEN

Dataset Collection

VizWiz (VizWiz, Gurari et al.)



33k Image + Question Pairs

from blind people



10 Short Answer per Question

from crowd workers

VizWiz-LF

Sampled 60

Original VizWiz
~1.7 words

vs

Ours
~30 words

Identification

Description

Reading

Misc.

Human Experts

GPT-4V

InstructBLIP

Gemini

LLaVA

BLIP-2

QWEN

Goal

Understanding Long-Form Answers to Visual Questions

RQ1. *What information is provided in long-form answers?*

RQ2. *How can we evaluate long-form answers?*

RQ3. *How can we mitigate hallucinations in long-form answers?*

What is the **role** of each sentence?

We identified **8 functional roles** in long-form answers

Confirmation

Answer

Auxiliary

Suggestion

Explanation

Answer Failure

Auxiliary Failure

Misc.

What is the **role** of each sentence?

Q. *What is in this bottle?*



GPT-4V

The image is blurry, but it shows a bottle with a label that includes the word MUSTARD. Given the color of the liquid visible through the bottle, it's likely to be a type of mustard sauce based on its golden-brown hue. Mustard sauces are commonly used as a condiment or dressing for sandwiches, salads, and many other dishes. If you have a clearer picture, I can provide a more accurate identification.

What is the **role** of each sentence?

Q. *What is in this bottle?*



GPT-4V

The image is blurry, but it shows a bottle with a label that includes the word MUSTARD. Given the color of the liquid visible through the bottle, it's likely to be a type of mustard sauce based on its golden-brown hue. Mustard sauces are commonly used as a condiment or dressing for sandwiches, salads, and many other dishes. If you have a clearer picture, I can provide a more accurate identification.

**Confirmation
of Photo**

What is the **role** of each sentence?

Q. *What is in this bottle?*



GPT-4V

The image is blurry, but it shows a bottle with a label that includes the word MUSTARD. Given the color of the liquid visible through the bottle, it's likely to be a type of mustard sauce based on its golden-brown hue. Mustard sauces are commonly used as a condiment or dressing for sandwiches, salads, and many other dishes. If you have a clearer picture, I can provide a more accurate identification.

Explanation

Answer

What is the **role** of each sentence?

Q. *What is in this bottle?*



GPT-4V

The image is blurry, but it shows a bottle with a label that includes the word MUSTARD. Given the color of the liquid visible through the bottle, it's likely to be a type of mustard sauce based on its golden-brown hue. Mustard sauces are commonly used as a condiment or dressing for sandwiches, salads, and many other dishes. If you have a clearer picture, I can provide a more accurate identification.

Auxiliary Information

What is the **role** of each sentence?

Q. *What is in this bottle?*



GPT-4V

The image is blurry, but it shows a bottle with a label that includes the word MUSTARD. Given the color of the liquid visible through the bottle, it's likely to be a type of mustard sauce based on its golden-brown hue. Mustard sauces are commonly used as a condiment or dressing for sandwiches, salads, and many other dishes. If you have a clearer picture, I can provide a more accurate identification.

Suggestion

What is the **role** of each sentence?

Q. *What is in this bottle?*



Expert

I can't tell what is in the bottle. If I had to guess, I would say it was some sort of cooking oil because it has an amber color and pictures of vegetables on the label. It could also be apple juice or another type of juice.

**Answer
Failure**

What is the **information source** of each sentence?

We identified **3 information types** in long-form answers

What is the **information source** of each sentence?

We identified **3 information types** in long-form answers

Image Content

It shows a bottle with a label that includes the word MUSTARD.

What is the **information source** of each sentence?

We identified **3 information types** in long-form answers

Image Content

It shows a bottle with a label that includes the word MUSTARD.

Image Quality

The image is quite blurry and the details are not clear.

What is the **information source** of each sentence?

We identified **3 information types** in long-form answers

Image Content

It shows a bottle with a label that includes the word MUSTARD.

Image Quality

The image is quite blurry and the details are not clear.

External

Mustard sauces are commonly used as a dressing for sandwiches, salads, and many other dishes.

Annotation and Classification

Open coding to
derive the taxonomy

8 Functional Roles

3 Information Types

Annotation and Classification

Open coding to
derive the taxonomy

8 Functional Roles

3 Information Types



N=180

Three-way 
human annotations

Fleis Kappa = 0.76

Fleis Kappa = 0.87

Annotation and Classification

Open coding to derive the taxonomy

8 Functional Roles

3 Information Types



N=180

Three-way  **human annotations**

Fleis Kappa = 0.76

Fleis Kappa = 0.87



N=4.2k

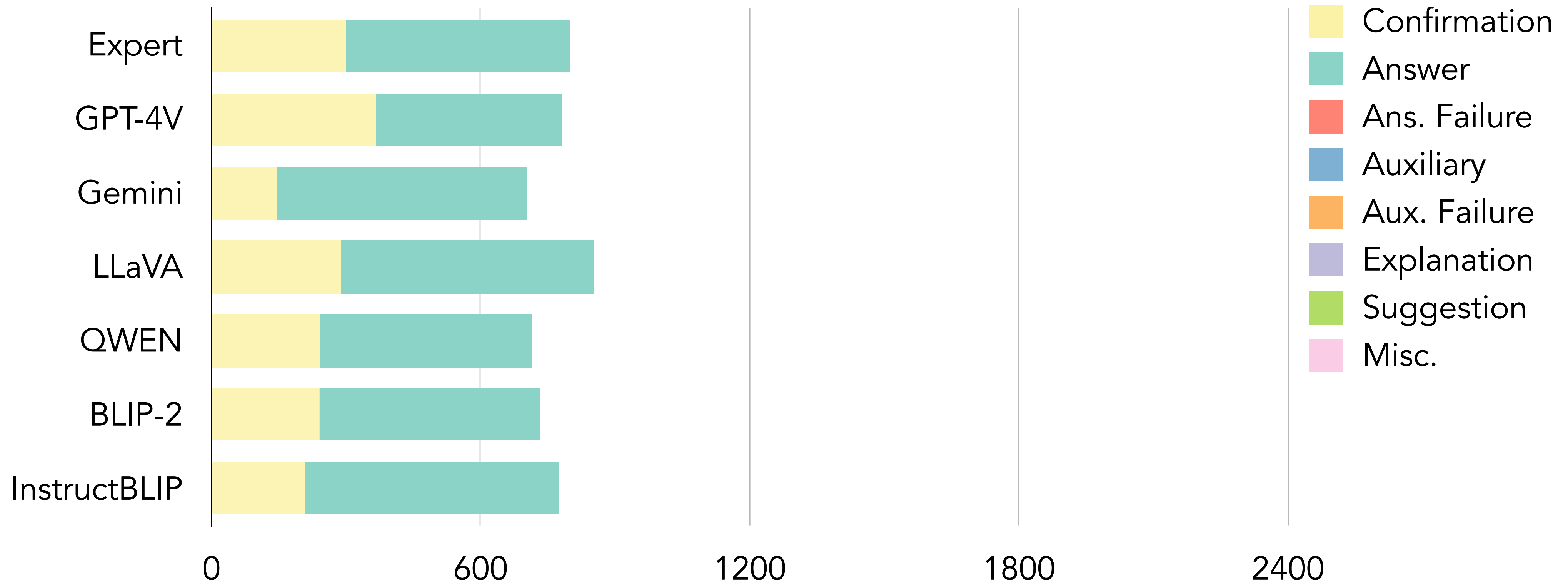
GPT-4 classifier w/ 8-shot examples

Fleis Kappa = 0.72

Fleis Kappa = 0.85

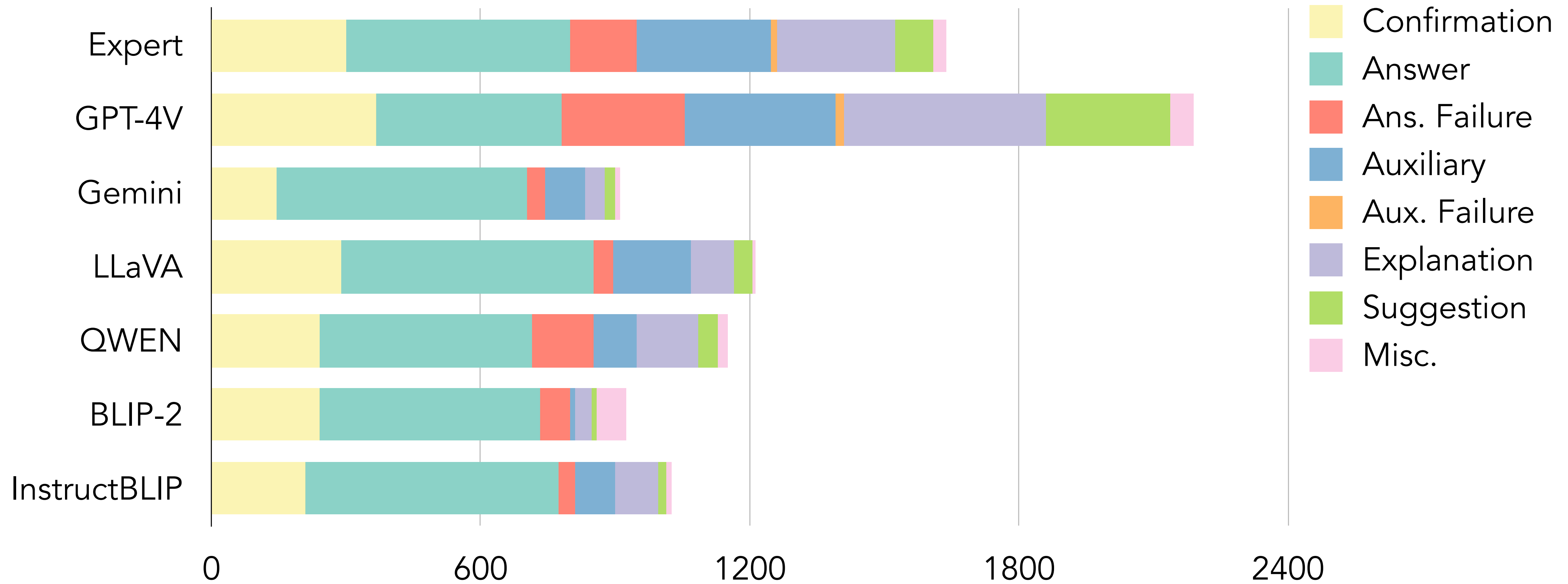
Results

Answers with each functional role



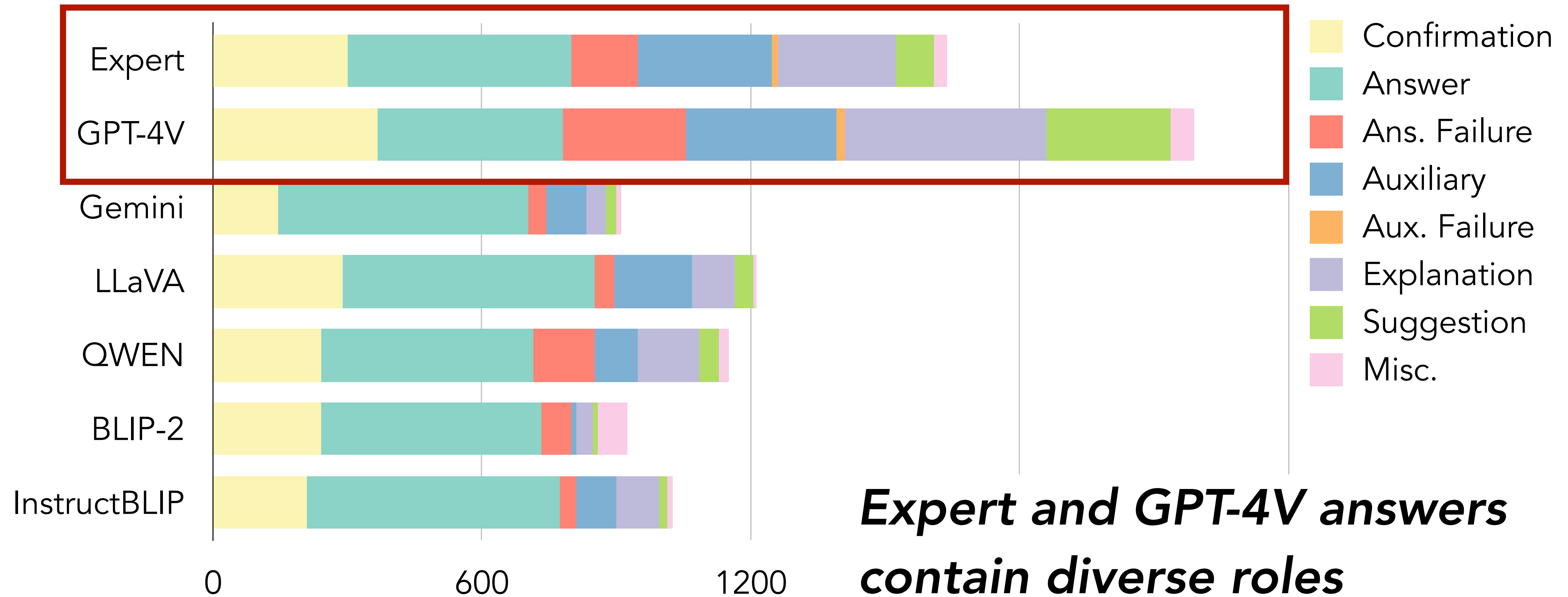
Results

Answers with each functional role



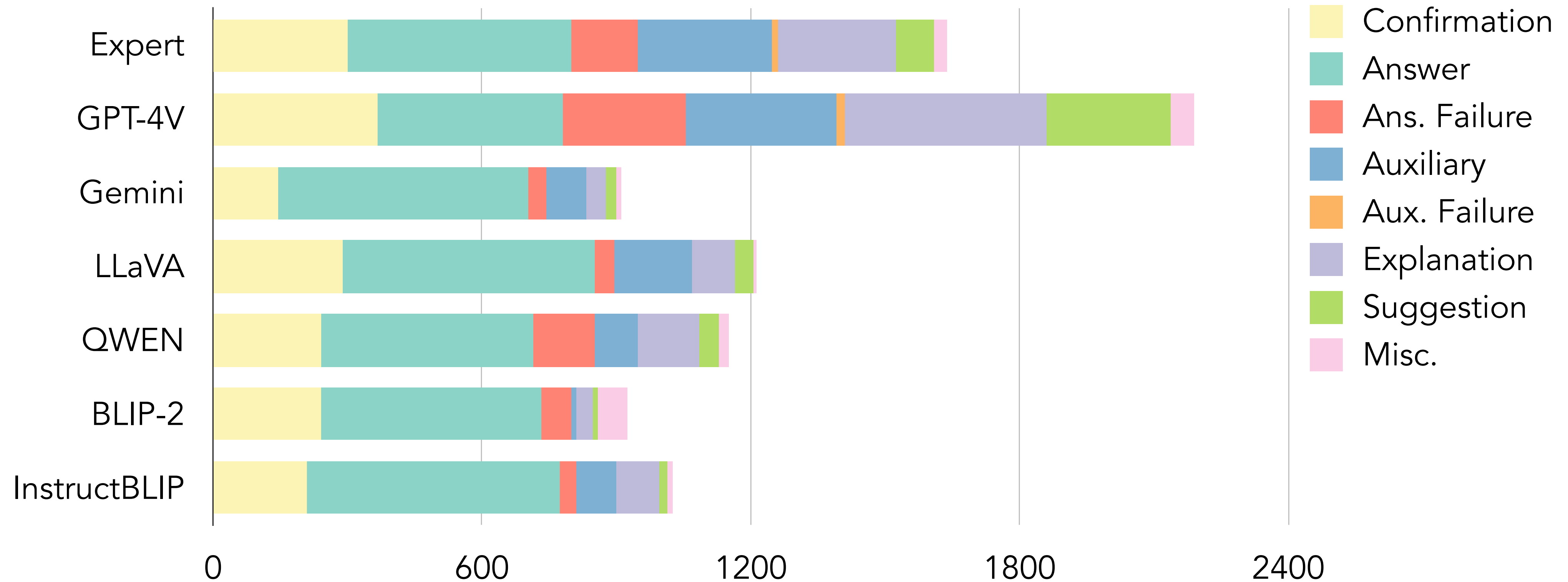
Results

Answers with each functional role



Results

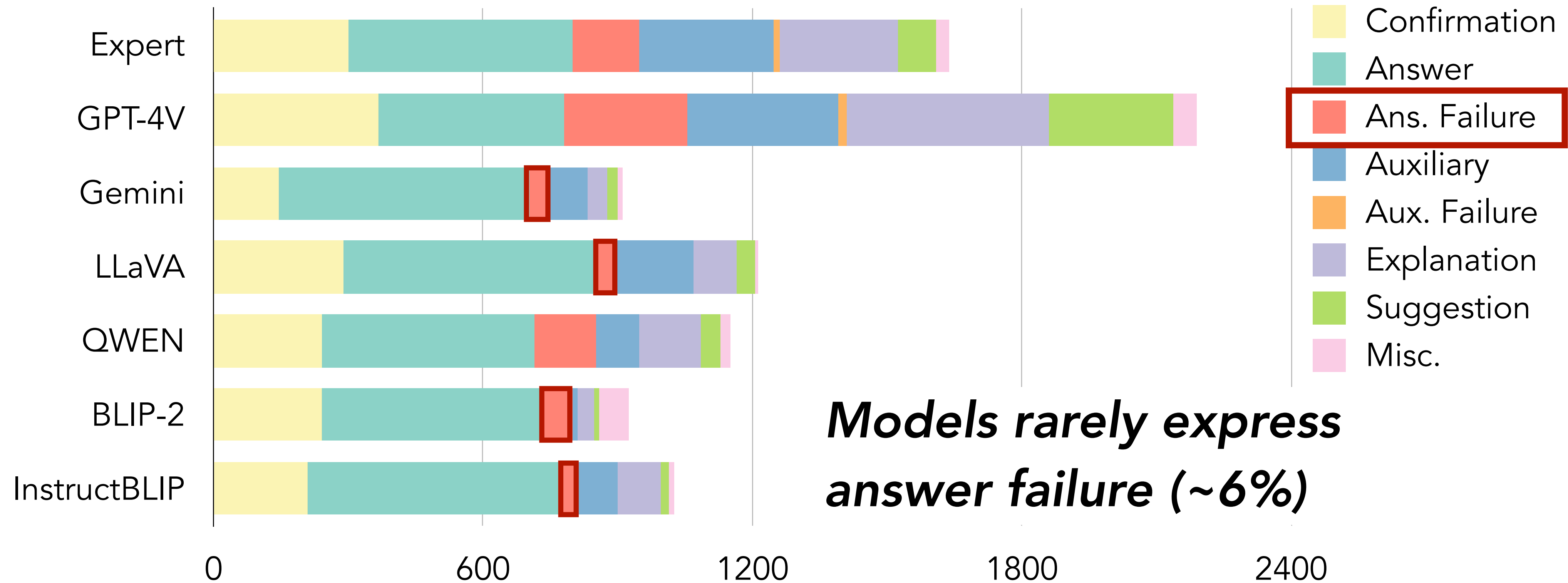
Answers with each functional role



*40% of questions marked unanswerable by humans

Results

Answers with each functional role

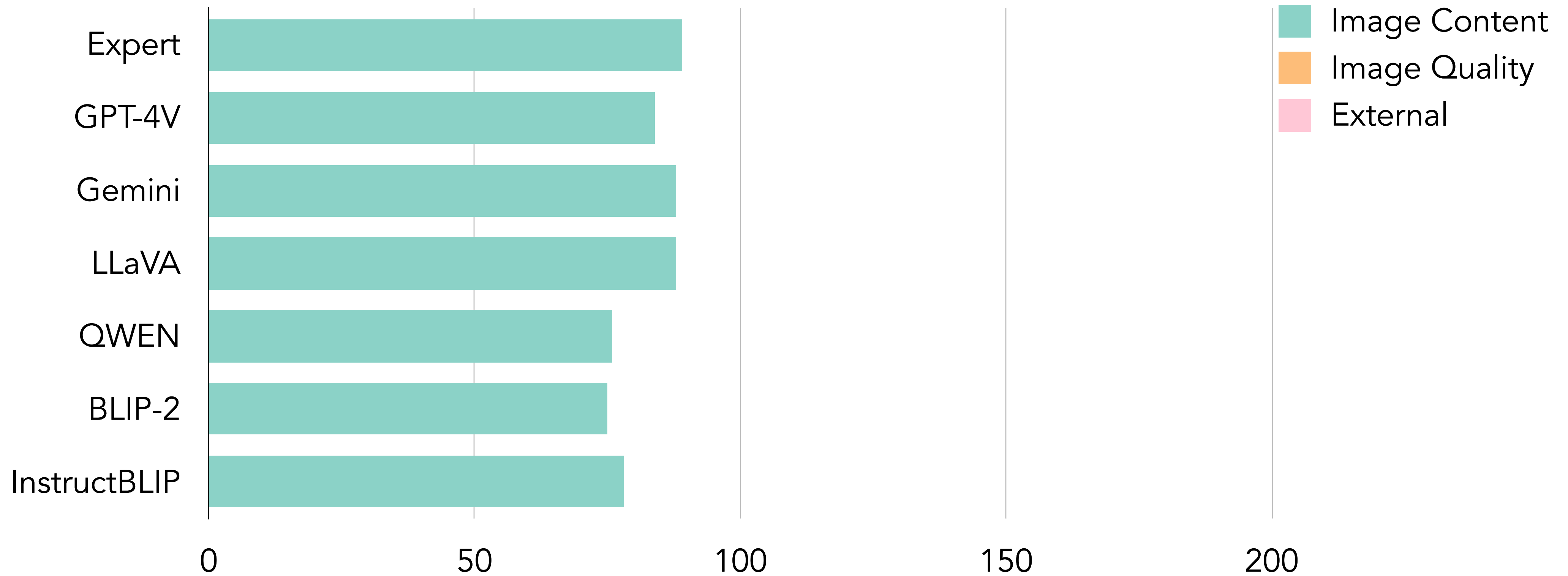


Models rarely express answer failure (~6%)

*40% of questions marked unanswerable by humans

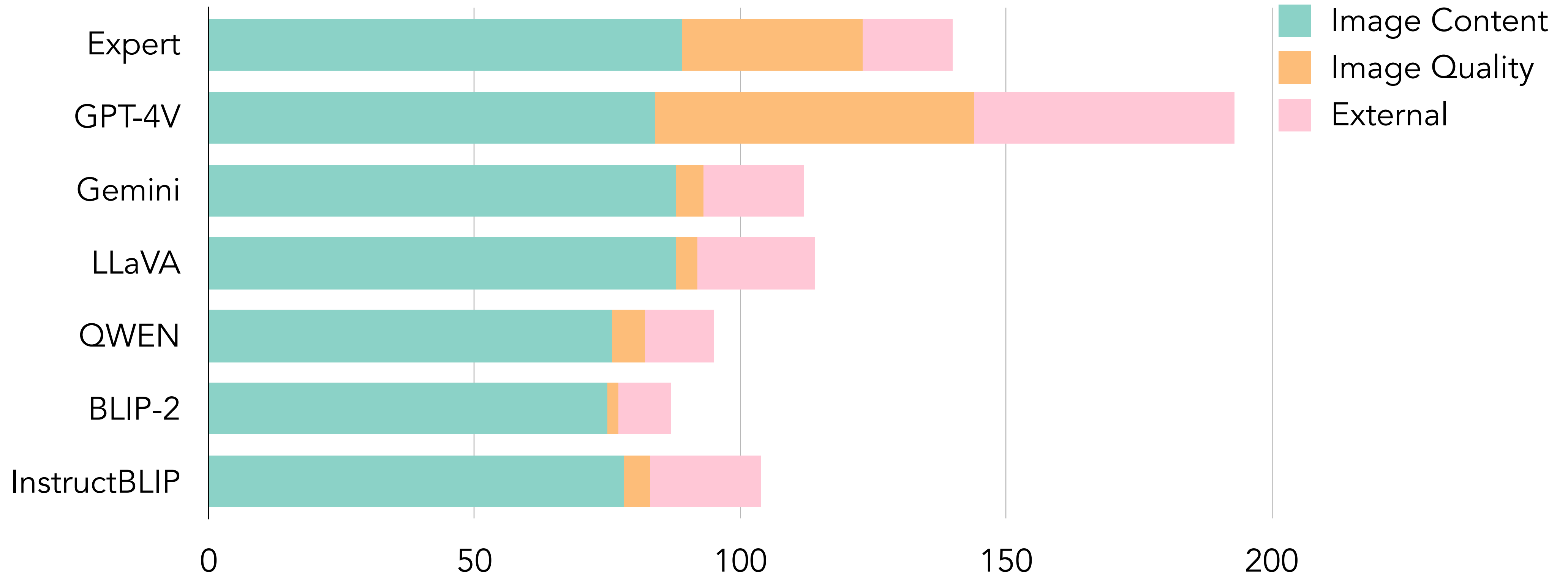
Results

Answers with each information type



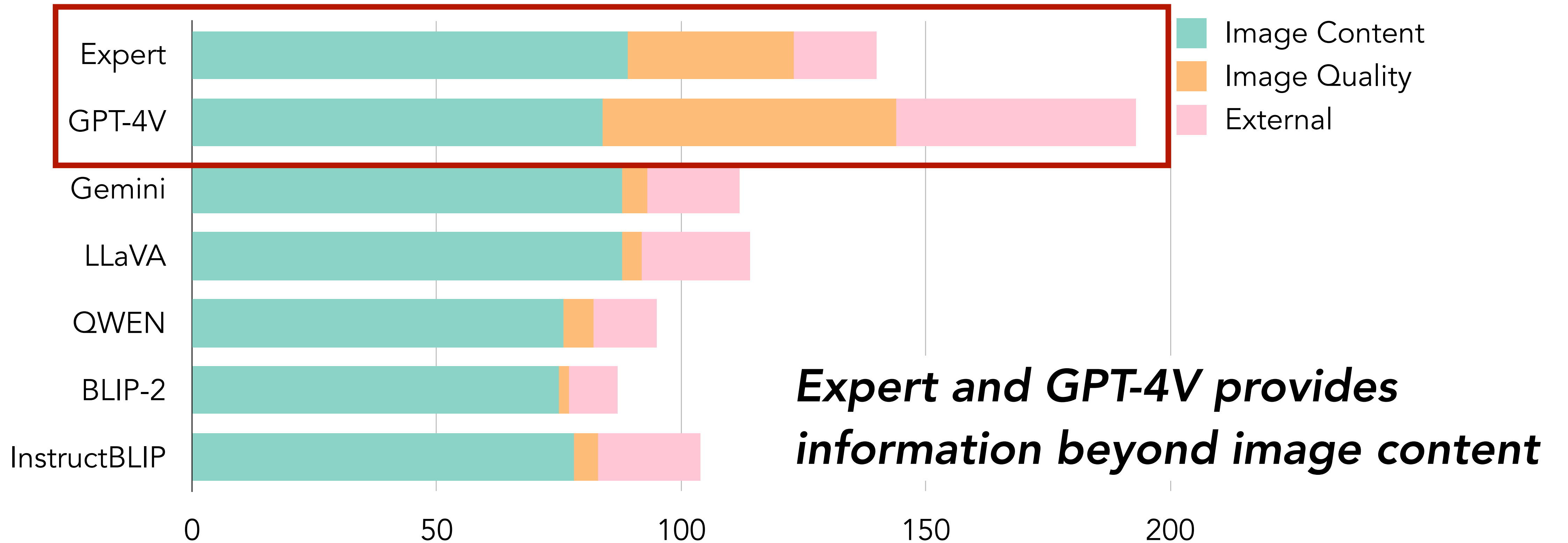
Results

Answers with each information type



Results

Answers with each information type



Goal

Understanding Long-Form Answers to Visual Questions

RQ1. *What information is provided in long-form answers?*

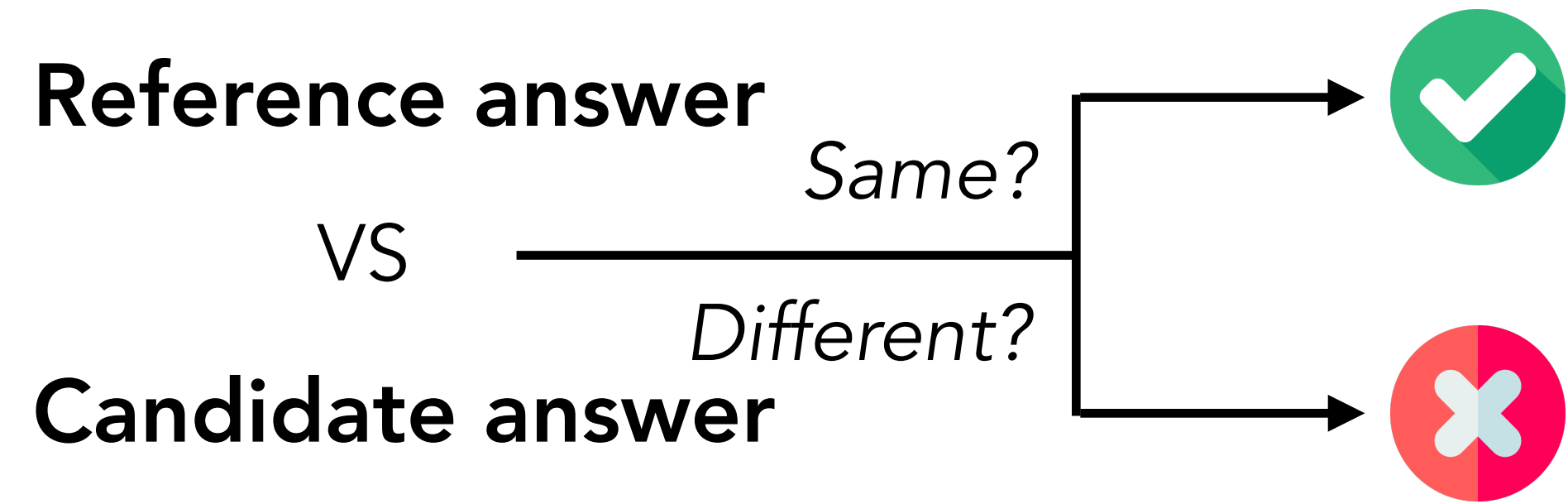
RQ2. *How can we evaluate long-form answers?*

RQ3. *How can we mitigate hallucinations in long-form answers?*

How can we automatically evaluate long-form answers?

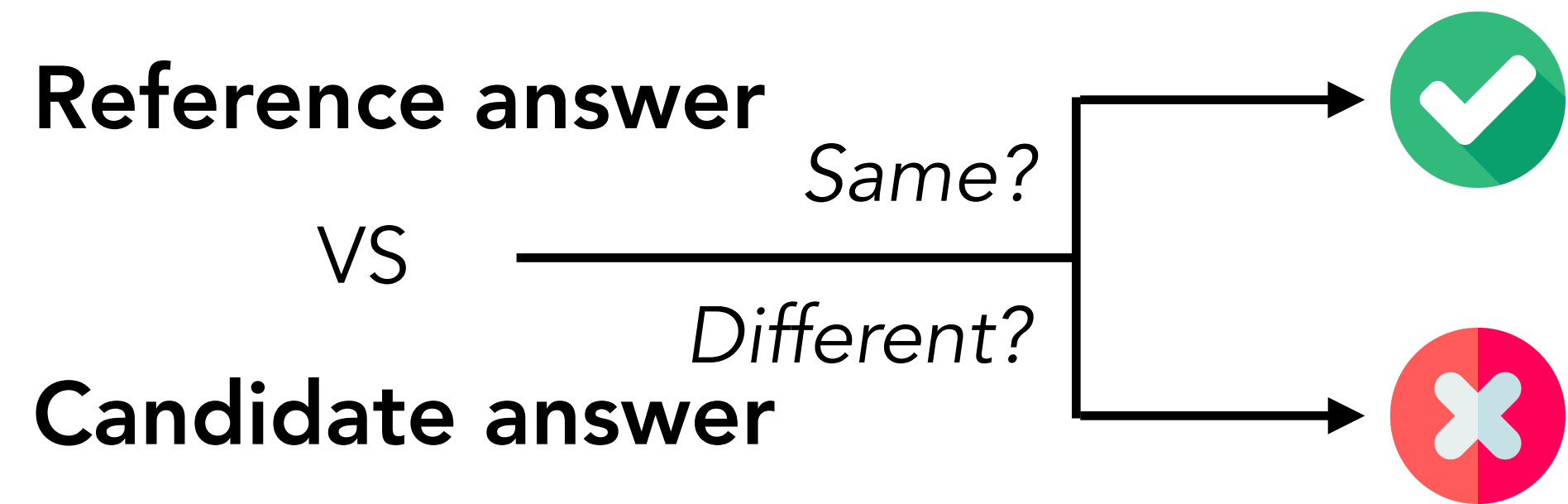
How can we automatically evaluate long-form answers?

Reference-based metrics



How can we automatically evaluate long-form answers?

Reference-based metrics



Traditional

ROUGE

METEOR

BERTScore

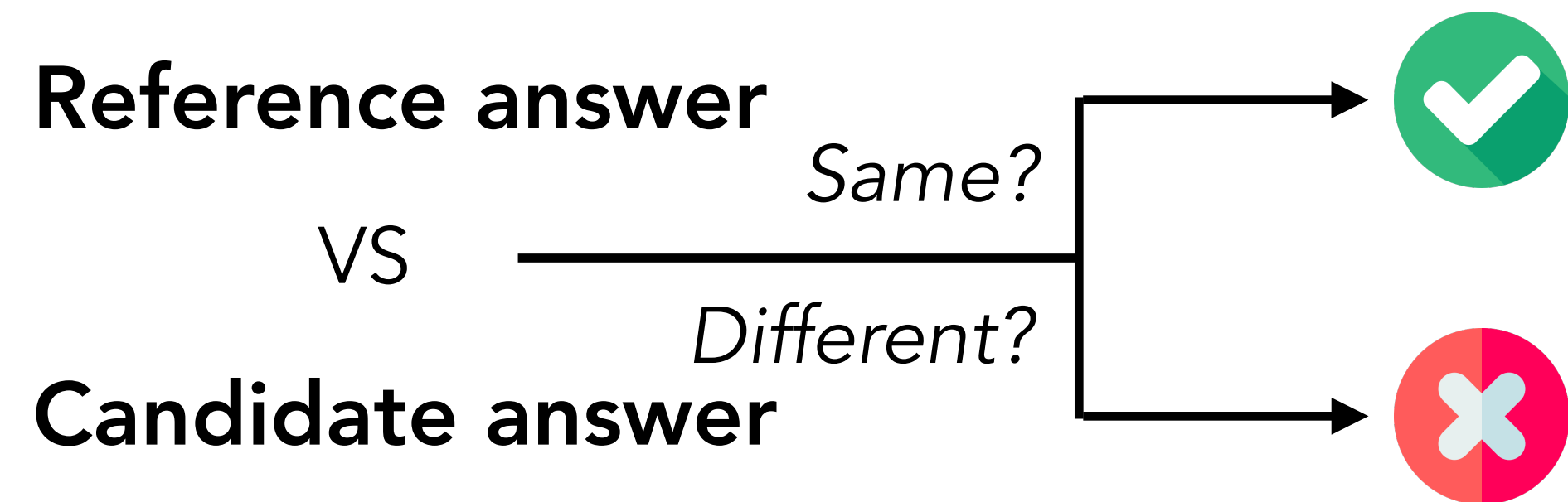
LLM-based

GPT-4

(LAVE, Mañas et al.)

How can we automatically evaluate long-form answers?

Reference-based metrics



Short Reference?
Long Reference?

Traditional

ROUGE

METEOR

BERTScore

LLM-based

GPT-4

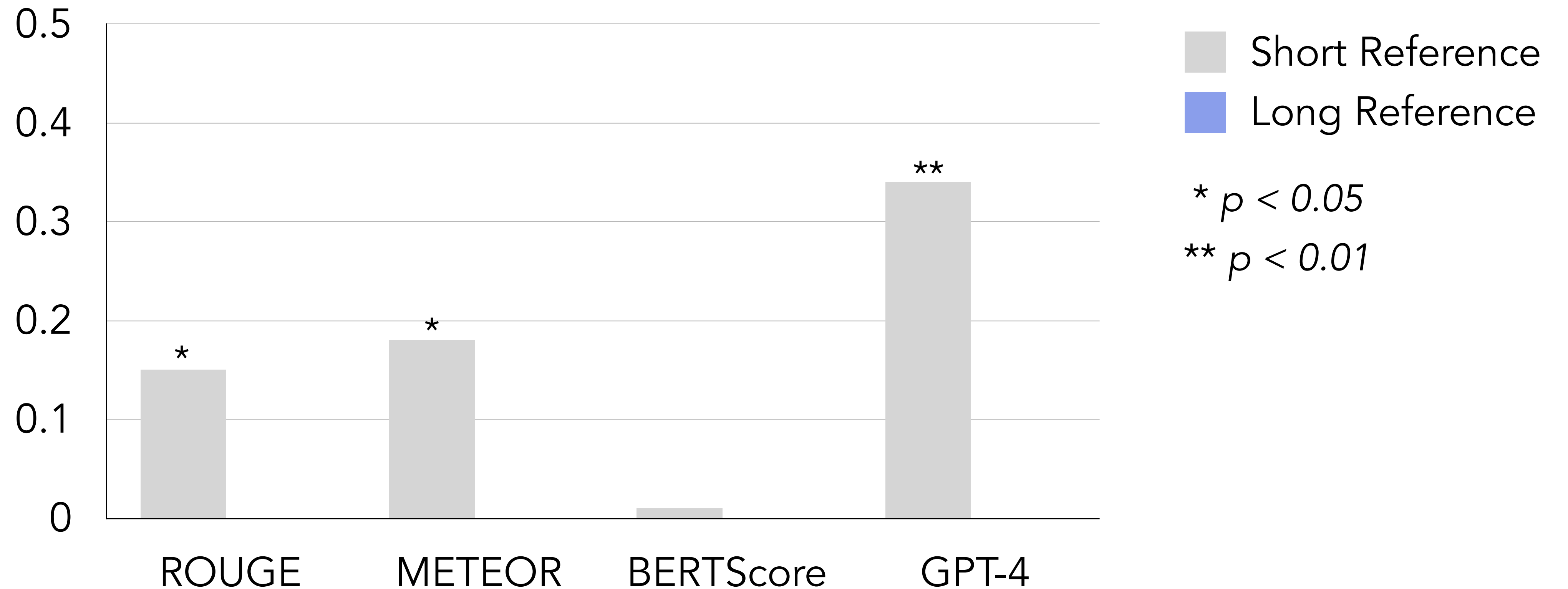
(LAVE, Mañas et al.)

How can we automatically evaluate long-form answers?

Pearson Correlation with Human ratings

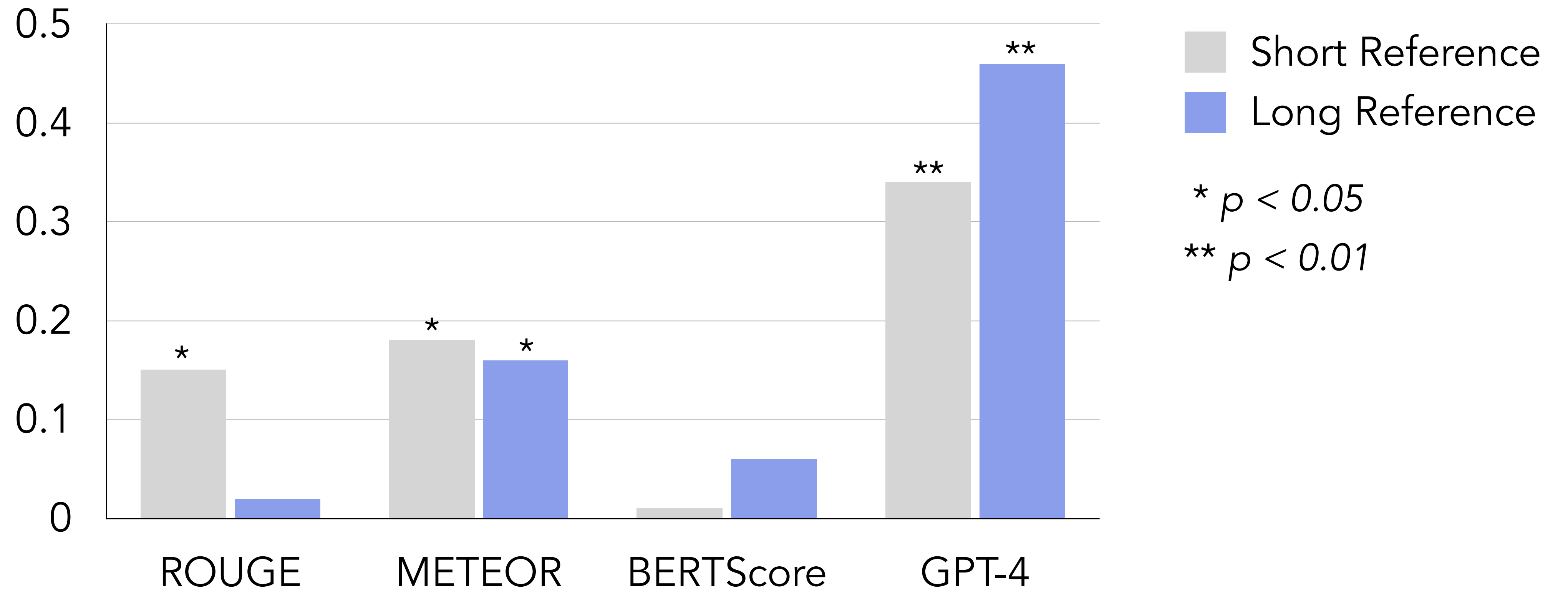
How can we automatically evaluate long-form answers?

Pearson Correlation with Human ratings



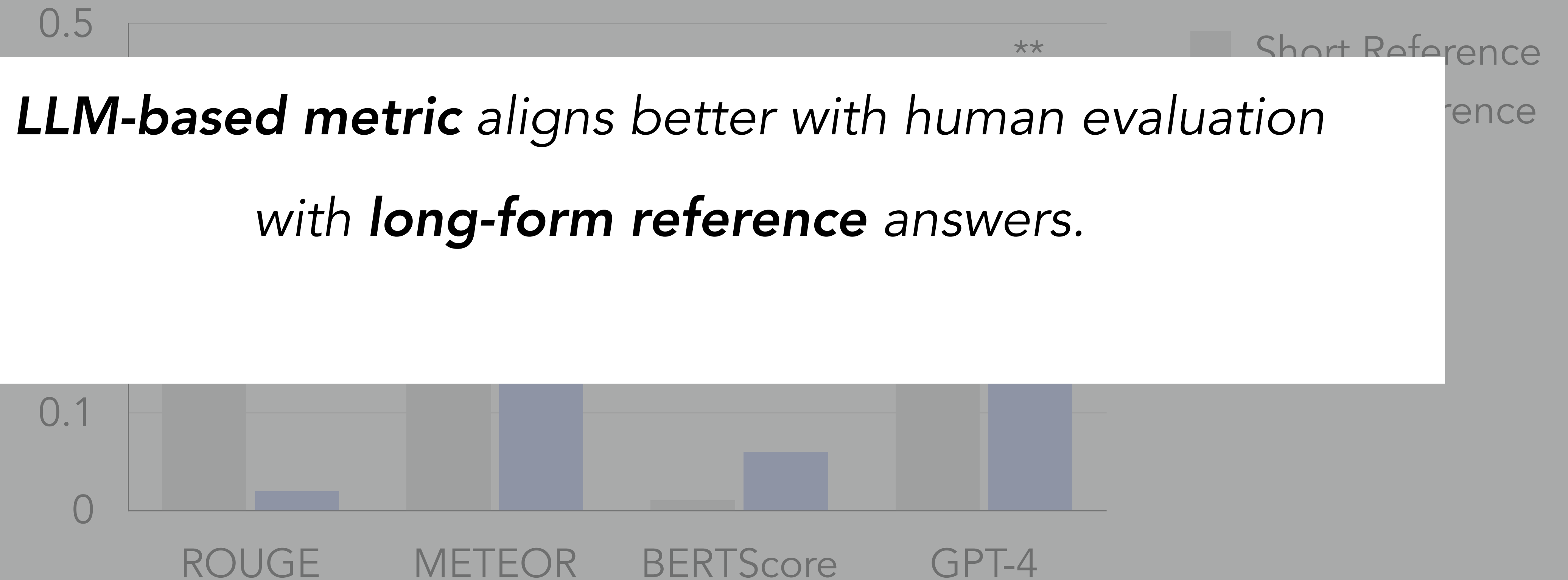
How can we automatically evaluate long-form answers?

Pearson Correlation with Human ratings



How can we automatically evaluate long-form answers?

Pearson Correlation with Human ratings



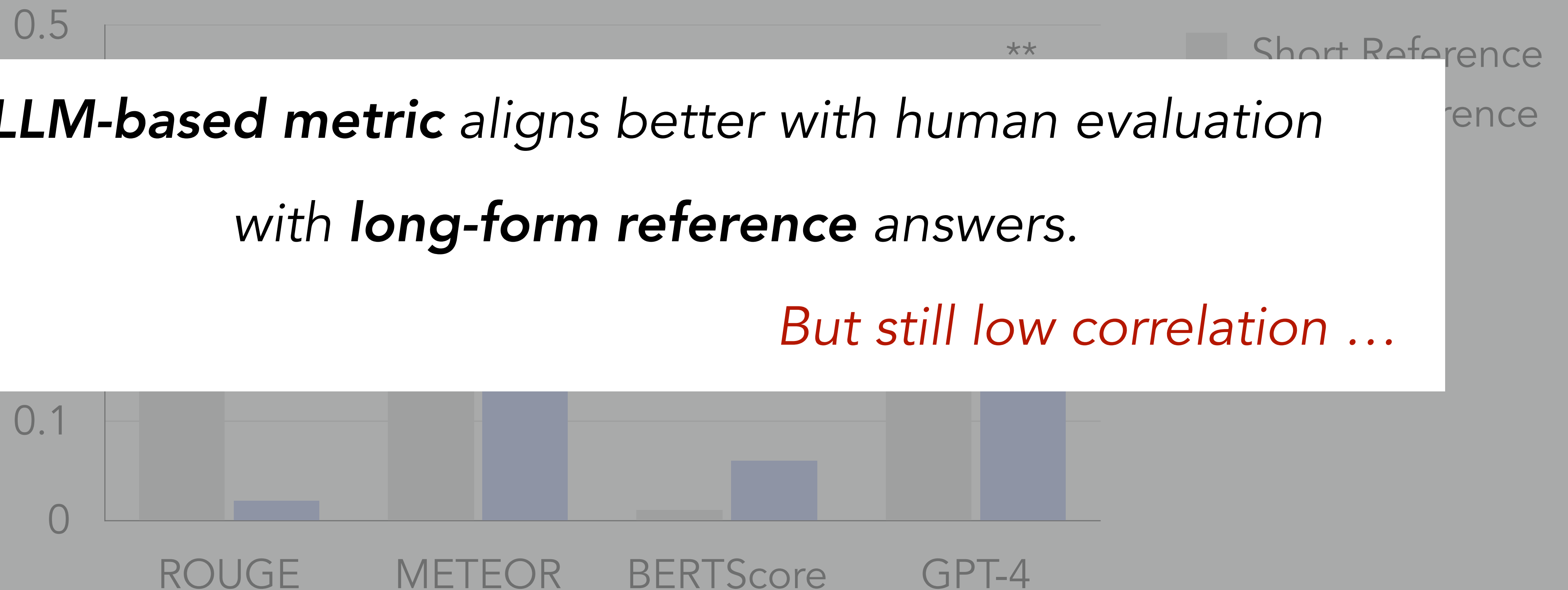
LLM-based metric aligns better with human evaluation with long-form reference answers.

How can we automatically evaluate long-form answers?

Pearson Correlation with Human ratings

LLM-based metric aligns better with human evaluation with long-form reference answers.

But still low correlation ...



How do humans evaluate long-form answers?

How do humans evaluate long-form answers?



- **40 Evaluators** (20 Blind, 20 Sighted)
- **36 Answers** (GPT-4V, Gemini, Expert)

How do humans evaluate long-form answers?



Sighted
with Image

Sighted
without Image

Blind &
Low Vision

- **40 Evaluators** (20 Blind, 20 Sighted)
- **36 Answers** (GPT-4V, Gemini, Expert)

How do humans evaluate long-form answers?



- **40 Evaluators** (20 Blind, 20 Sighted)
- **36 Answers** (GPT-4V, Gemini, Expert)

Sighted
with Image

Sighted
without Image

Blind &
Low Vision

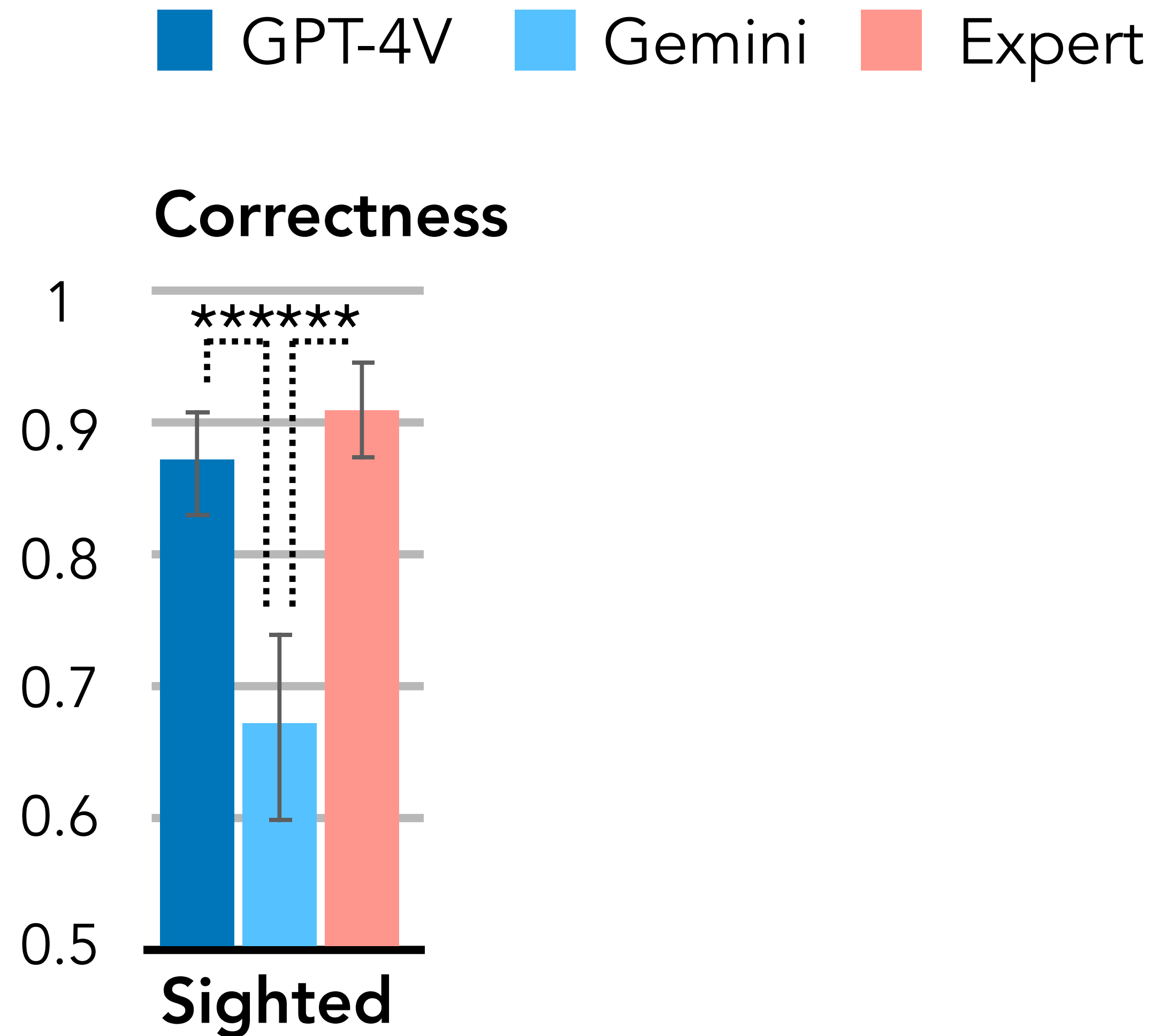
Correctness

Plausibility

How do humans evaluate long-form answers?



- **40 Evaluators** (20 Blind, 20 Sighted)
- **36 Answers** (GPT-4V, Gemini, Expert)

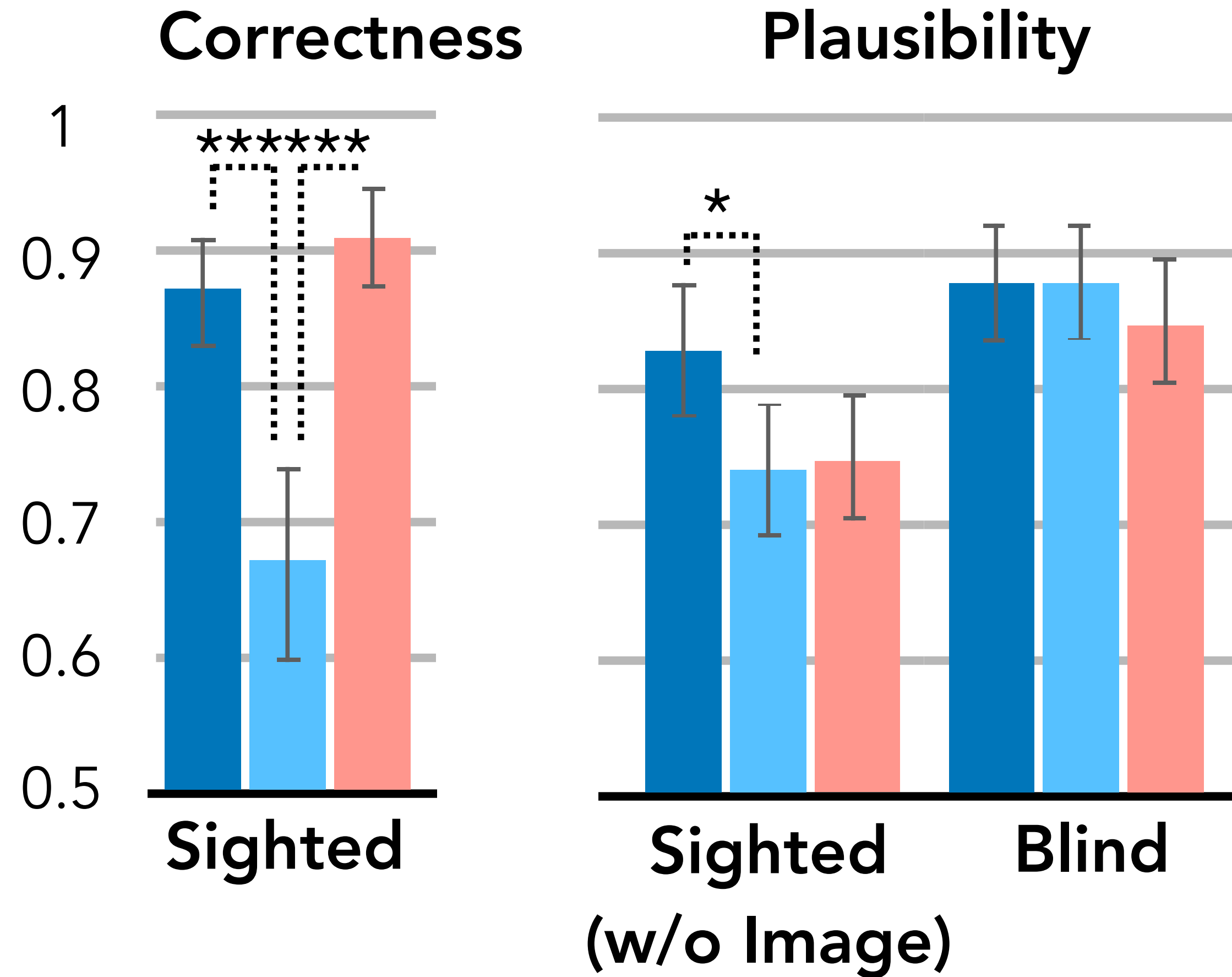


How do humans evaluate long-form answers?



- **40 Evaluators** (20 Blind, 20 Sighted)
- **36 Answers** (GPT-4V, Gemini, Expert)

■ GPT-4V ■ Gemini ■ Expert



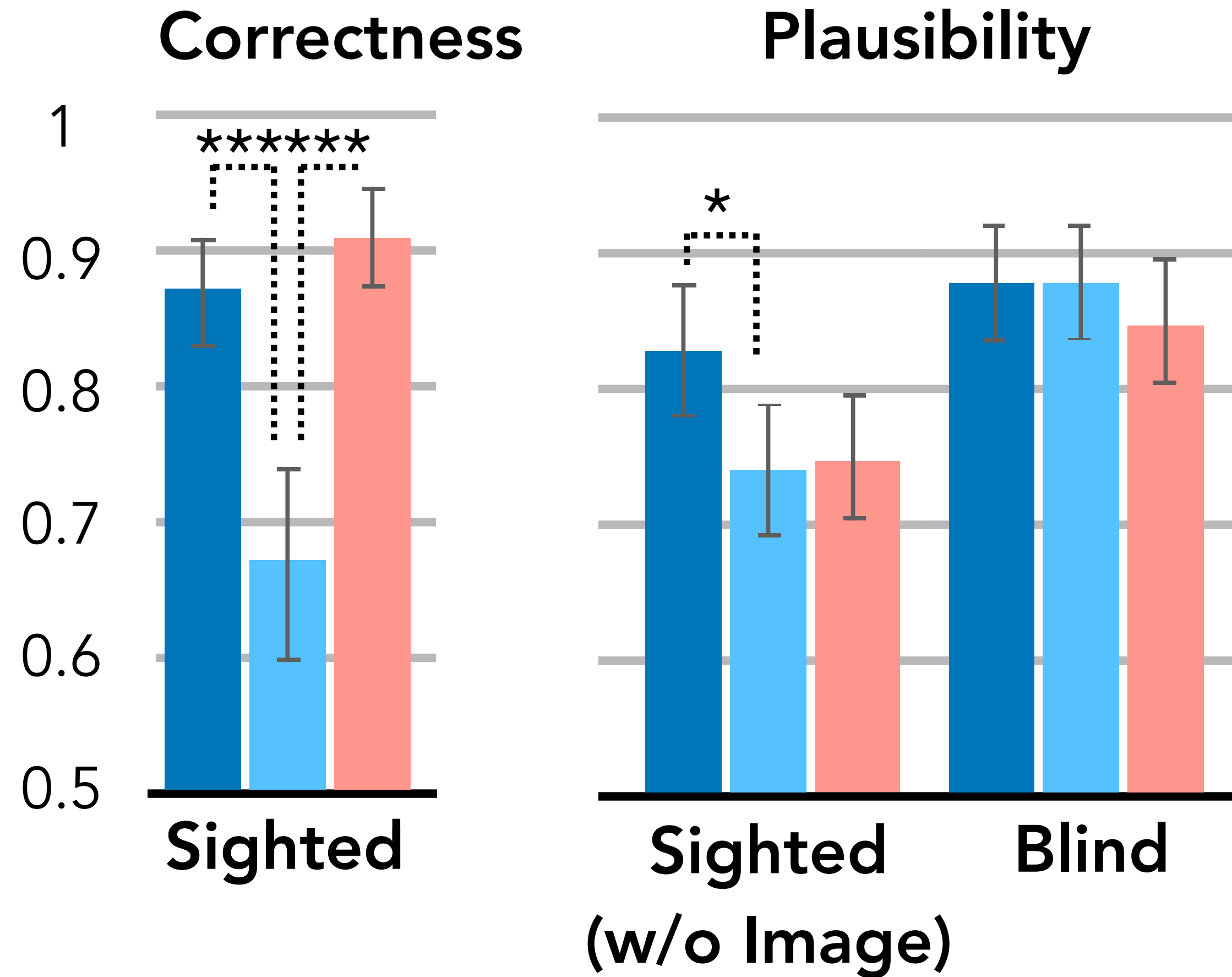
How do humans evaluate long-form answers?



- 40 Evaluators (20 Blind, 20 Sighted)
- 36 Answers (GPT-4V, Gemini, Expert)

When do models generate incorrect answers?

■ GPT-4V ■ Gemini ■ Expert



Models rarely abstain to **unanswerable** questions

Q. *What tablet size is it?*



Models rarely abstain to **unanswerable** questions

Q. *What tablet size is it?*



Expert

This picture is too blurry and dark to make out details like writing. Can you send another photo with brighter lighting conditions and a little further away? [...]

GPT-4V

The tablet The image you've provided is of low quality and quite blurry, making it difficult to discern any specific details, including the tablet size. [...]

**Answer
Failure**

Models rarely abstain to **unanswerable** questions

Q. *What tablet size is it?*



Gemini

It's hard to tell exactly what size the tablet is, but it looks like it might be a size 00 capsule.

LLaVA

The tablet size in the image is a small tablet.

QWEN

The tablet size is large.

BLIP-2

a 7-inch tablet with 128gb of internal storage and an external micro-sd card slot

InstructBLIP

10 x 20 cms 2 inches by 1 inch in size and weight

Models rarely abstain to **unanswerable** questions

Q. *What tablet size is it?*



Gemini

It's hard to tell exactly what size the tablet is, but it looks like it might be a size 00 capsule.

LLaVA

The tablet size in the image is a small tablet.

QWEN

The tablet size is large.

BLIP-2

a 7-inch tablet with 128gb of internal storage and an external micro-sd card slot

InstructBLIP

10 x 20 cms 2 inches by 1 inch in size and weight

These models hallucinate with "answer" sentences!

Goal

Understanding Long-Form Answers to Visual Questions

RQ1. *What information is provided in long-form answers?*

RQ2. *How can we evaluate long-form answers?*

RQ3. *How can we mitigate hallucinations in long-form answers?*

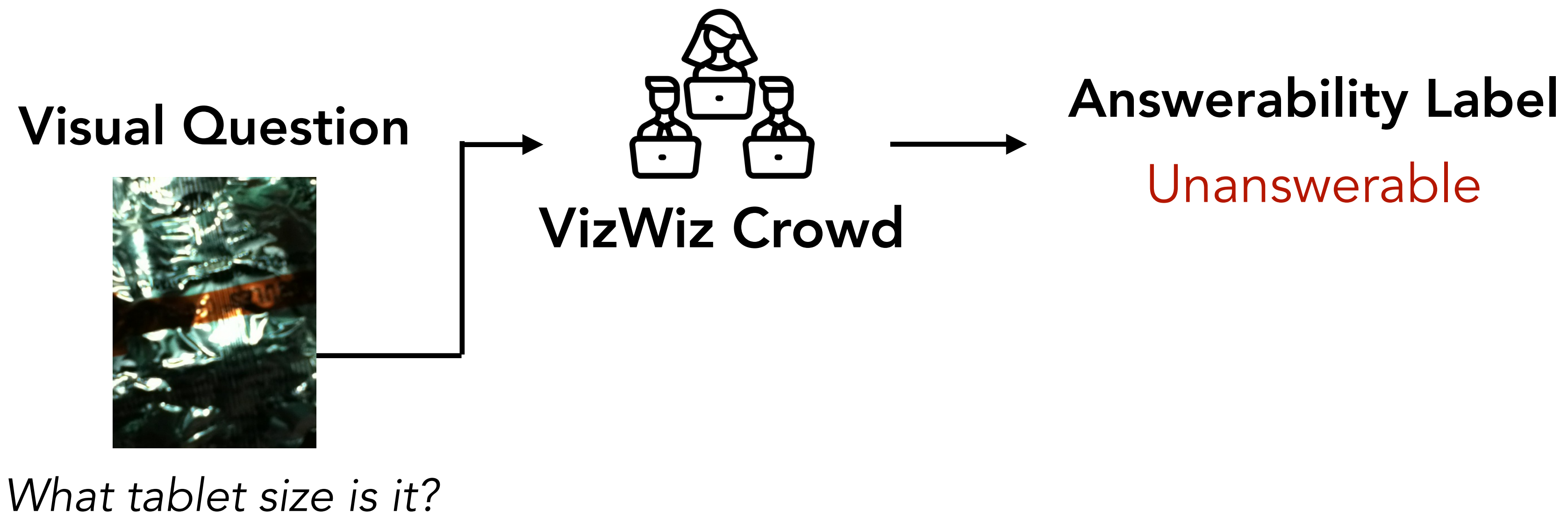
Evaluating if a model correctly abstained

Visual Question

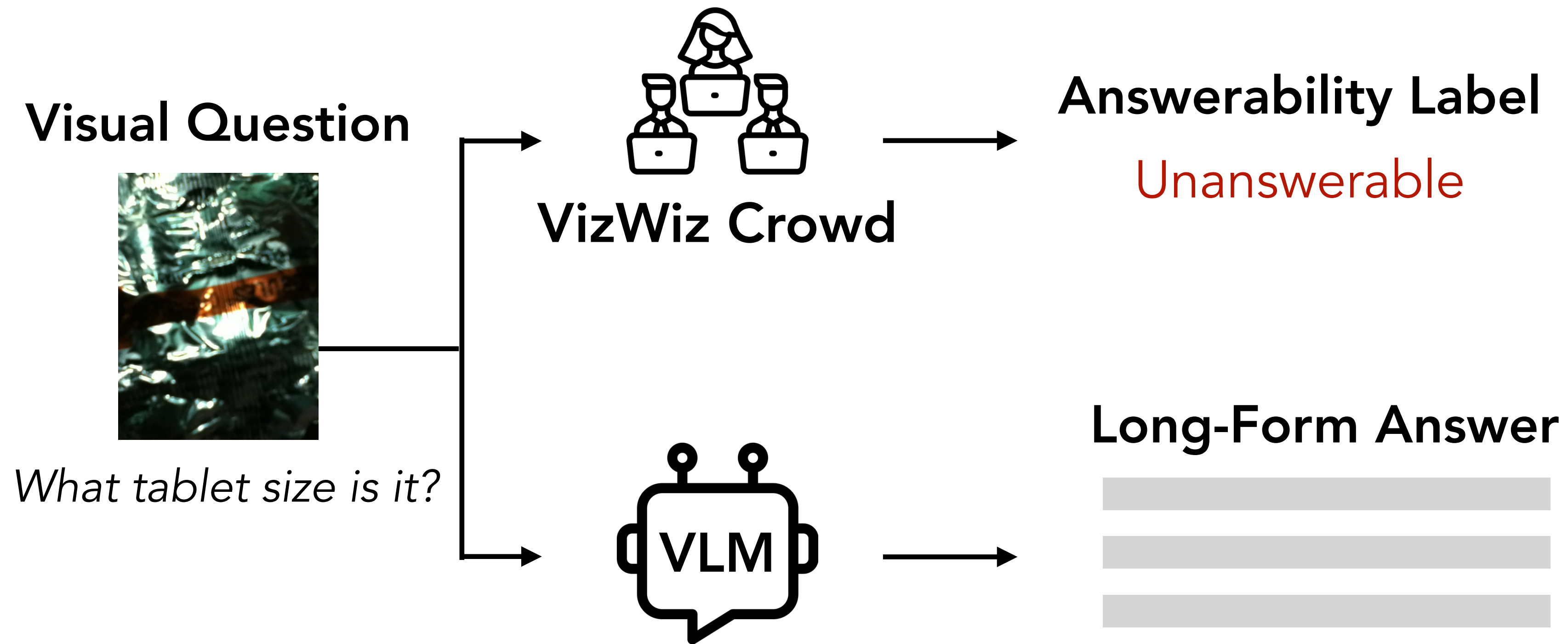


What tablet size is it?

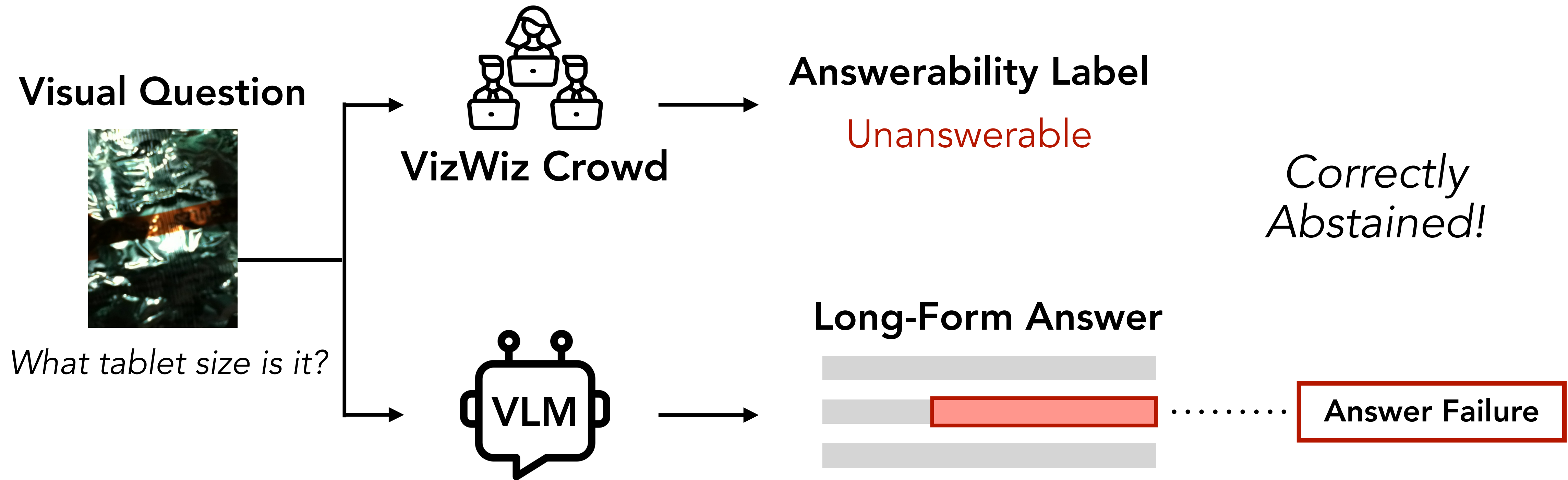
Evaluating if a model correctly abstained



Evaluating if a model correctly abstained



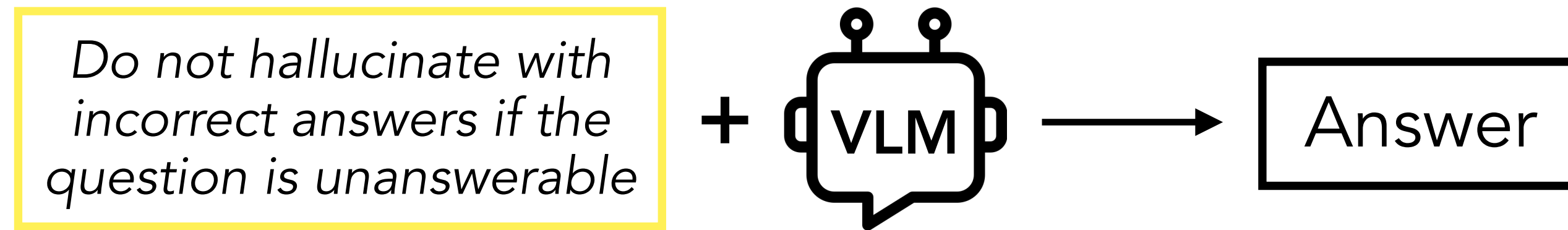
Evaluating if a model correctly abstained



Abstention Approaches

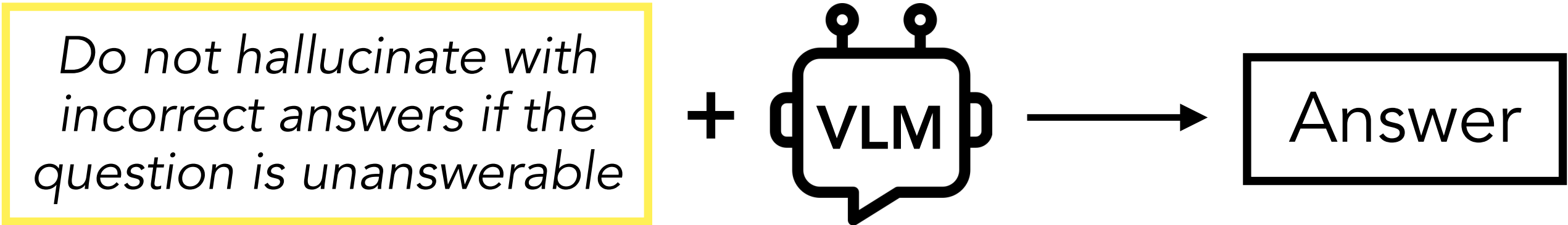
Abstention Approaches

Abstention Instruction

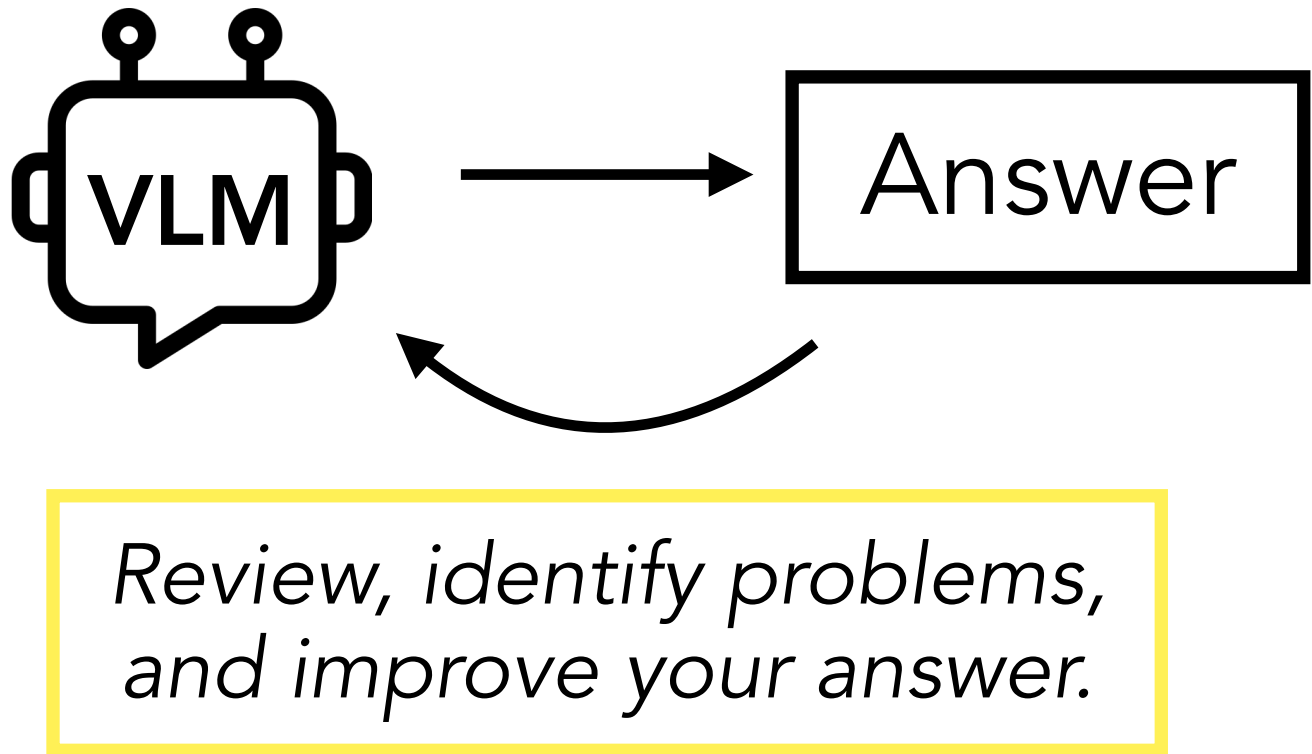


Abstention Approaches

Abstention Instruction

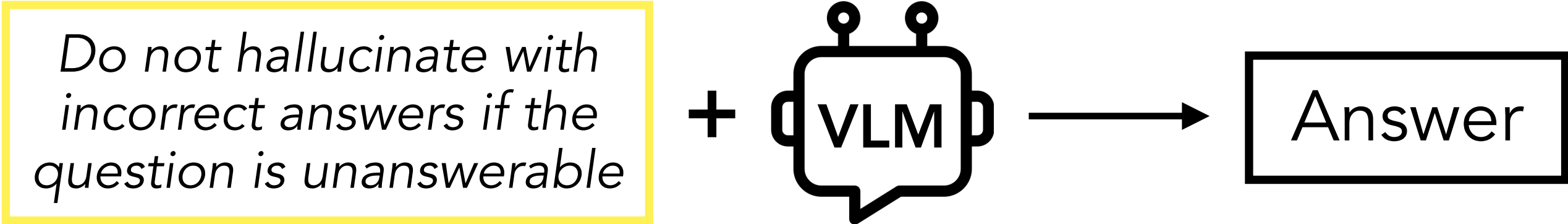


Self-Correction

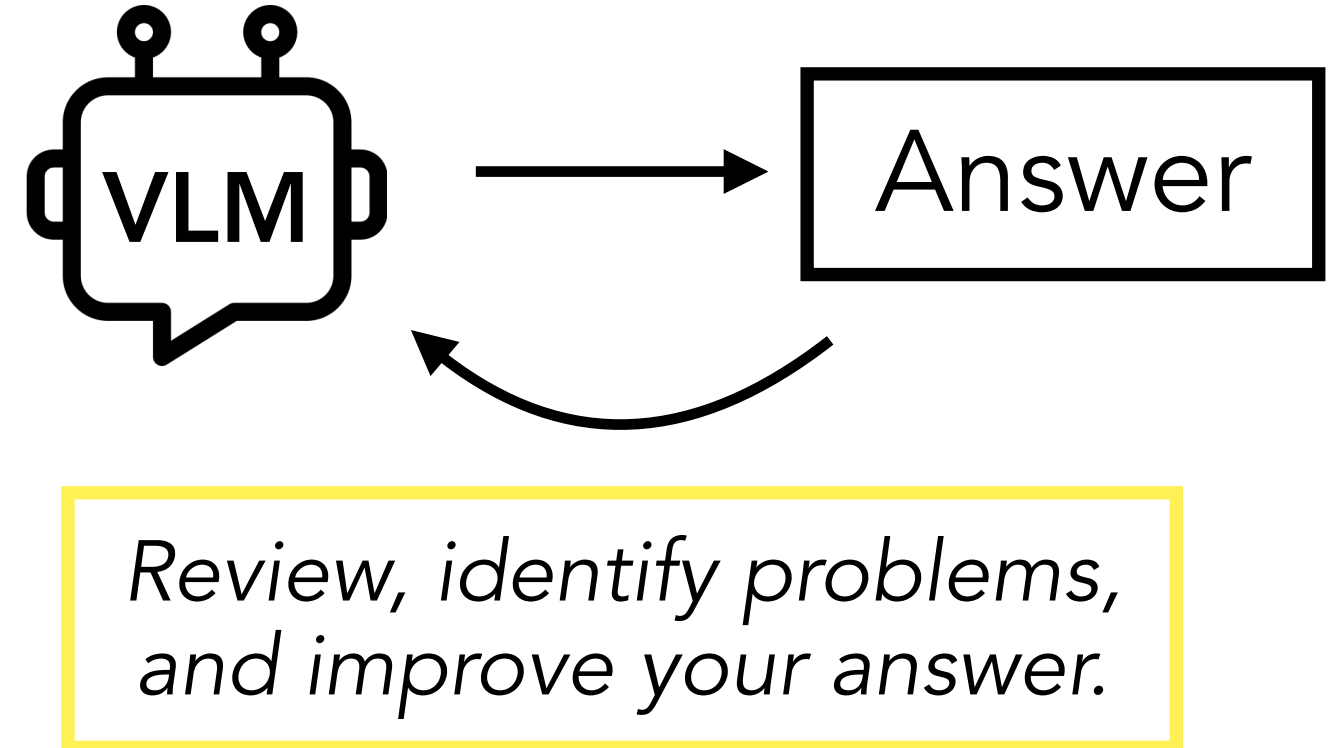


Abstention Approaches

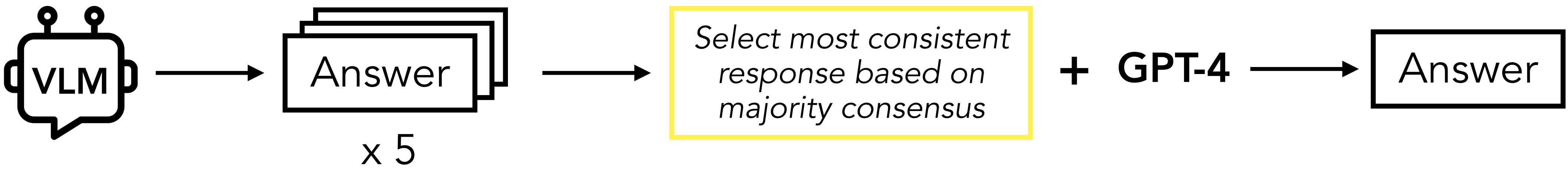
Abstention Instruction



Self-Correction

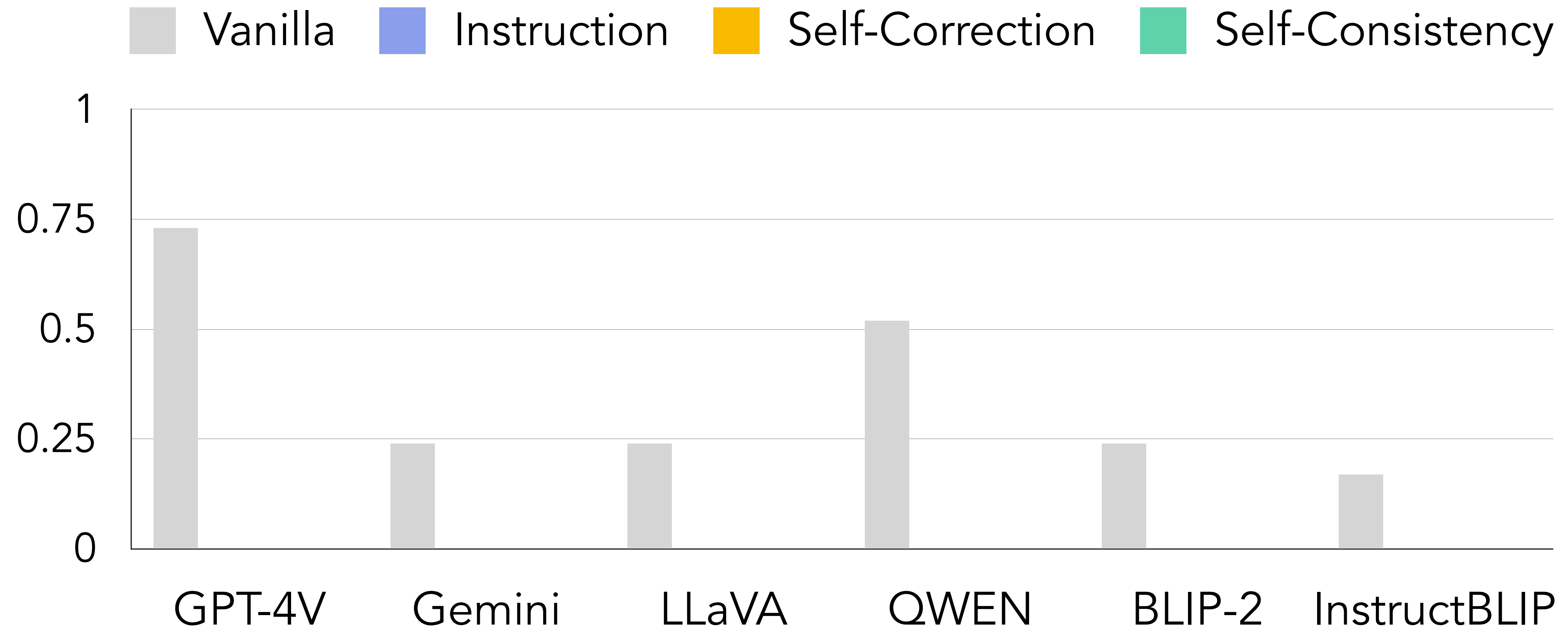


Self-Consistency



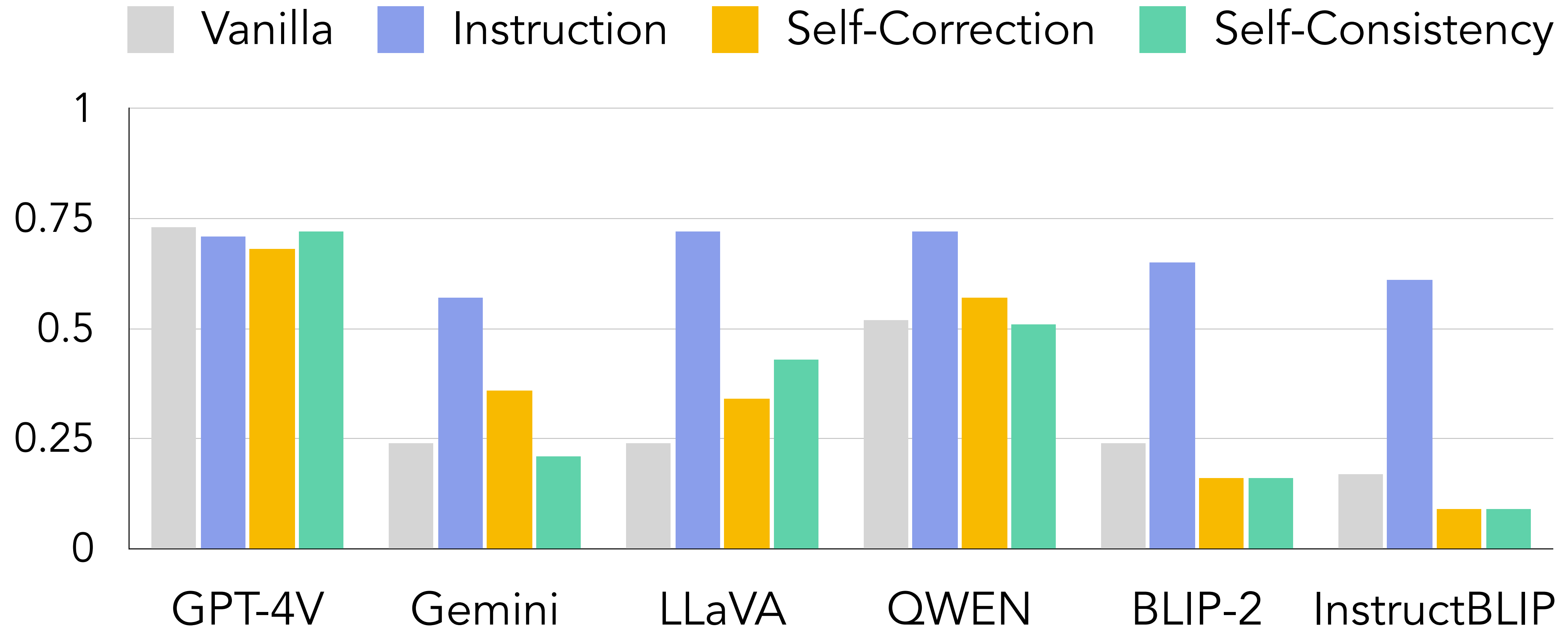
How can models abstain to unanswerable questions?

Abstention (F1)



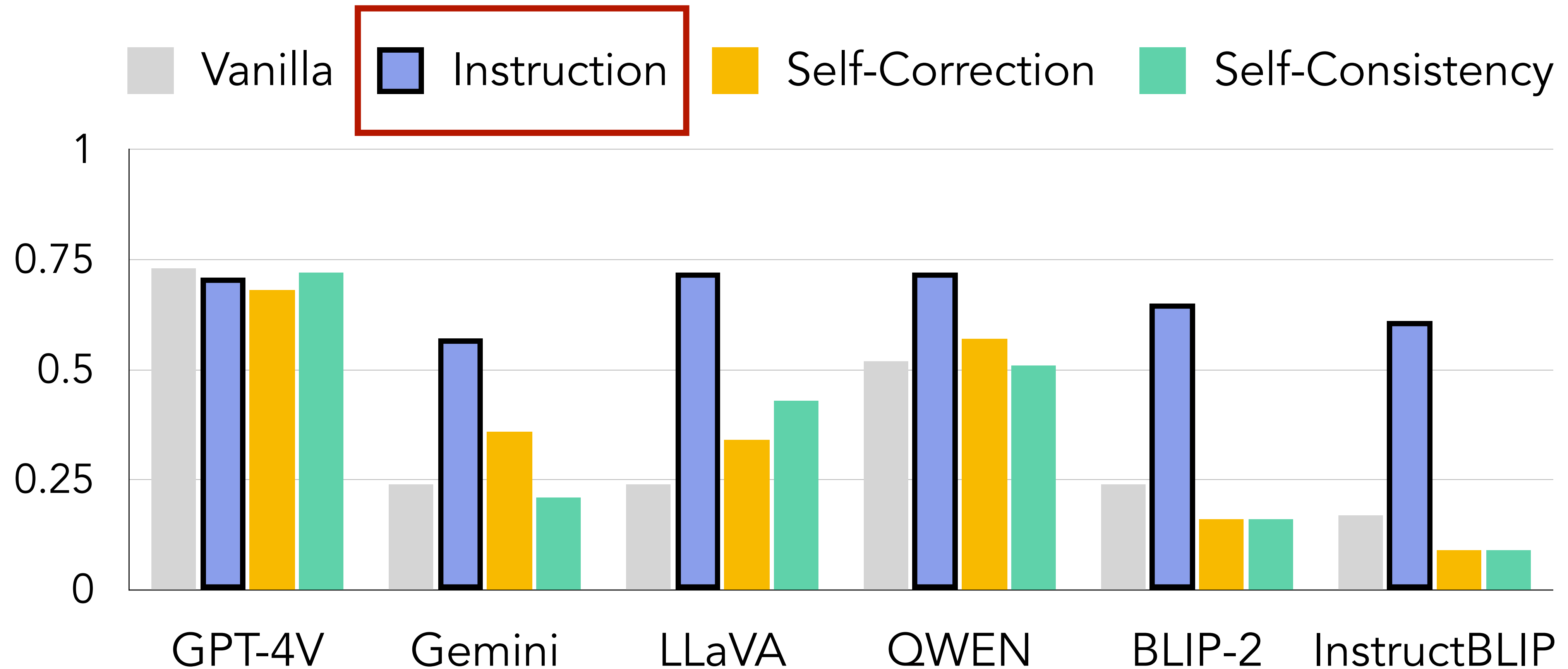
How can models abstain to unanswerable questions?

Abstention (F1)



How can models abstain to unanswerable questions?

Abstention (F1)



How can models abstain to unanswerable questions?

Abstention (F1)



Summary

RQ1. *What information is provided in long-form answers?*

Content Analysis

Image Quality Answer Supplementary

The image is rotated and partially obscured. From the visible label, it appears to be a can of Campbell's Condensed Soup, specifically the "Cream of Mushroom" variety. This type of soup contains a creamy mushroom broth and is often used as a base in various recipes or consumed on its own as a soup.

RQ2. *How can we evaluate long-form answers?*

Evaluation

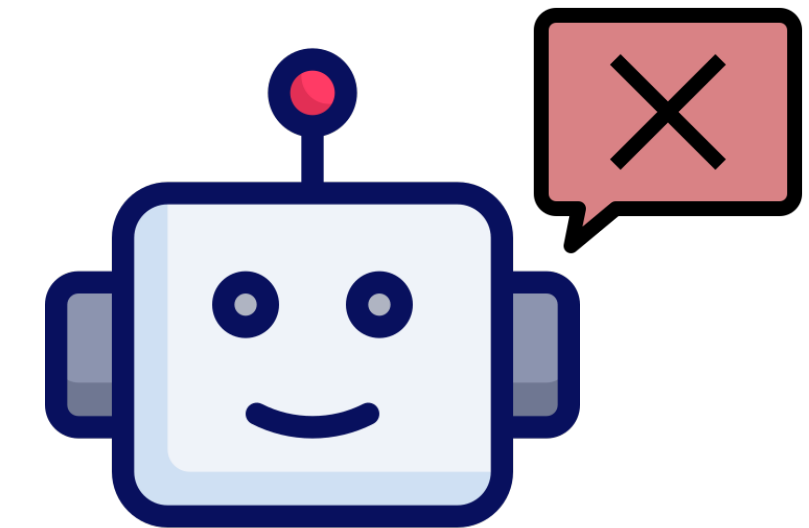
Automatic & Human Eval



RQ3. *How can we mitigate hallucinations in long-form answers?*

Hallucination Mitigation

Abstention Approaches



Takeaways & Future Work

- Long-form answers contain **diverse information beyond answer** including extra-visual information.
- It is **difficult to evaluate long-form answers** with short references.
- Current vision language models **hallucinate with unanswerable questions.**

Takeaways & Future Work

- Long-form answers contain **diverse information beyond answer** including extra-visual information.
- It is **difficult to evaluate long-form answers** with short references.
- Current vision language models **hallucinate with unanswerable questions.**
- Future work can **investigate broader types of errors.**
e.g., Not providing information present in the photo, misleading information
- We encourage more comprehensive **evaluation of VLMs with end-users.**
e.g., Evaluating relevance, redundancy, clarity of long-form answers

Backup Slides

How can we automatically evaluate long-form answers?

Reference-based metrics

Q. *What is this?*



Long Reference

This is a box of Pringles Original. The text on the top of the box reads 'Pringle Bells' and the box appears to be a special edition Christmas box with decorative symbols like a Santa's cap and red and white color theme to signify that. The box is on a table [...]

Long Candidate

This is a can of Pringles potato chips. The image on the can shows the Pringles logo and an illustration of the original flavored chips. Pringles are known for their unique packaging in a cylindrical can and their saddle-shaped chips that stack neatly inside.

How can we automatically evaluate long-form answers?

Reference-based metrics

Q. *What is this?*



Short Reference

Pringles

Long Candidate

This is a can of Pringles potato chips. The image on the can shows the Pringles logo and an illustration of the original flavored chips. Pringles are known for their unique packaging in a cylindrical can and their saddle-shaped chips that stack neatly inside.

How can we automatically evaluate long-form answers?

Reference-based metrics

Q. *What is this?*



Short Reference

Pringles

Extracted Candidate **Answer**

This is a can of Pringles potato chips. The image on the can shows the Pringles logo and an illustration of the original flavored chips. Pringles are known for their unique packaging in a cylindrical can and their saddle-shaped chips that stack neatly inside.

How can we automatically evaluate long-form answers?

Reference-based metrics

Q. *What is this?*



Long Reference

This is a box of Pringles Original. The text on the top of the box reads 'Pringle Bells' and the box appears to be a special edition Christmas box with decorative symbols like a Santa's cap and red and white color theme to signify that. The box is on a table [...]

Long Candidate

This is a can of Pringles potato chips. The image on the can shows the Pringles logo and an illustration of the original flavored chips. Pringles are known for their unique packaging in a cylindrical can and their saddle-shaped chips that stack neatly inside.

How can we automatically evaluate long-form answers?

Reference-based metrics

Q. *What is this?*



Long Reference

This is a box of Pringles Original. The text on the top of the box reads 'Pringle Bells' and the box appears to be a special edition Christmas box with decorative symbols like a Santa's cap and red and white color theme to signify that. The box is on a table [...]

Long Candidate

This is a can of Pringles potato chips. The image on the can shows the Pringles logo and an illustration of the original flavored chips. Pringles are known for their unique packaging in a cylindrical can and their saddle-shaped chips that stack neatly inside.

Human Eval
1.00 (Correct)

ROUGE
0.27

METEOR
0.15

BERTScore
0.87

LLM-based
1

How can we automatically evaluate long-form answers?

Reference-based metrics

Q. *What is this?*



Extracted Reference Answer

This is a box of Pringles Original. The text on the top of the box reads 'Pringle Bells' and the box appears to be a special edition Christmas box with decorative symbols like a Santa's cap and red and white color theme to signify that. The box is on a table [...]

Extracted Candidate Answer

This is a can of Pringles potato chips. The image on the can shows the Pringles logo and an illustration of the original flavored chips. Pringles are known for their unique packaging in a cylindrical can and their saddle-shaped chips that stack neatly inside.

Human Eval	ROUGE	METEOR	BERTScore	LLM-based
1.00 (Correct)	0.67	0.69	0.96	1

How can we automatically evaluate long-form answers?

Reference-based metrics

Q. *What is this?*



Extracted Reference Answer

This is a box of Pringles Original. The text on the top of the box reads 'Pringle Bells' and the box appears to be a special edition Christmas box with decorative symbols like a Santa's cap and red and white color theme to signify that. The box is on a table [...]

Extracted Candidate Answer

This is a can of Pringles potato chips. The image on the can shows the Pringles logo and an illustration of the original flavored chips. Pringles are known for their unique packaging in a cylindrical can and their saddle-shaped chips that stack neatly inside.

How can we automatically evaluate long-form answers?

Reference-based metrics

Q. *What is this?*



Extracted Reference Answer

This is a box of Pringles Original. The text on the top of the box reads 'Pringle Bells' and the box appears to be a special edition Christmas box with decorative symbols like a Santa's cap and red and white color theme to signify that. The box is on a table [...]

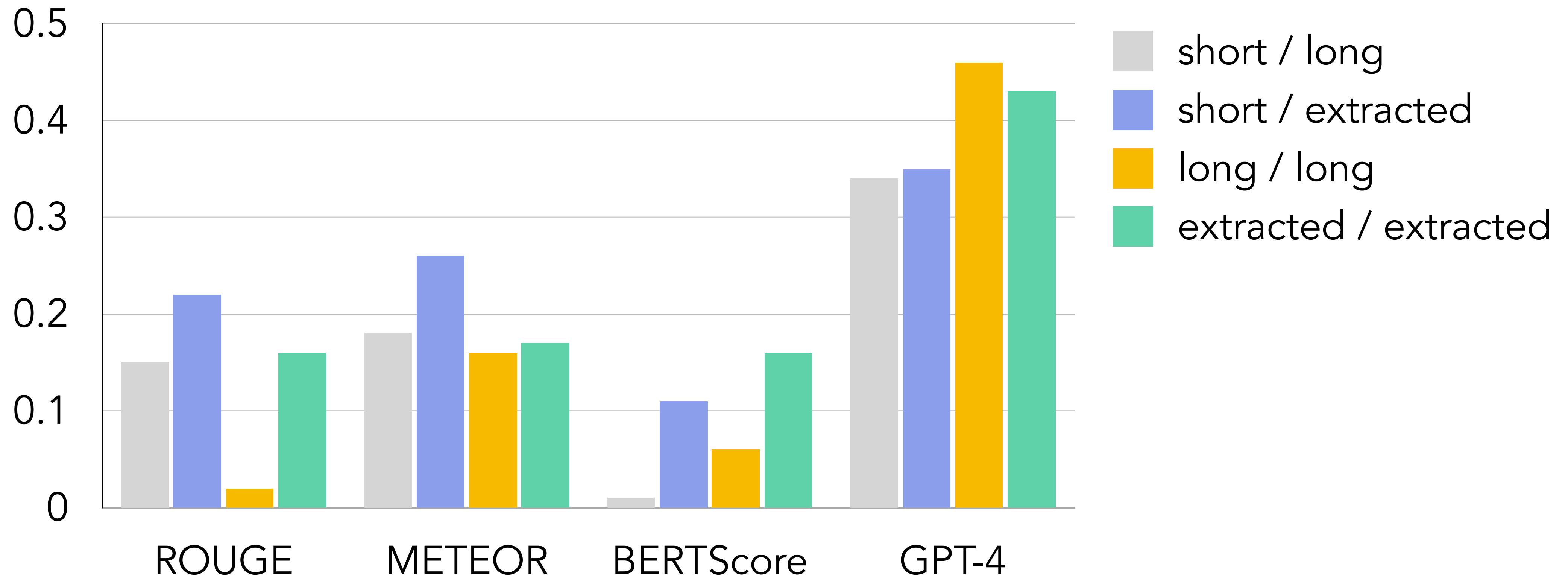
Extracted Candidate Answer

This is a can of Pringles potato chips. The image on the can shows the Pringles logo and an illustration of the original flavored chips. Pringles are known for their unique packaging in a cylindrical can and their saddle-shaped chips that stack neatly inside.

*We conducted **human evaluation** to evaluate beyond answer sentences.*

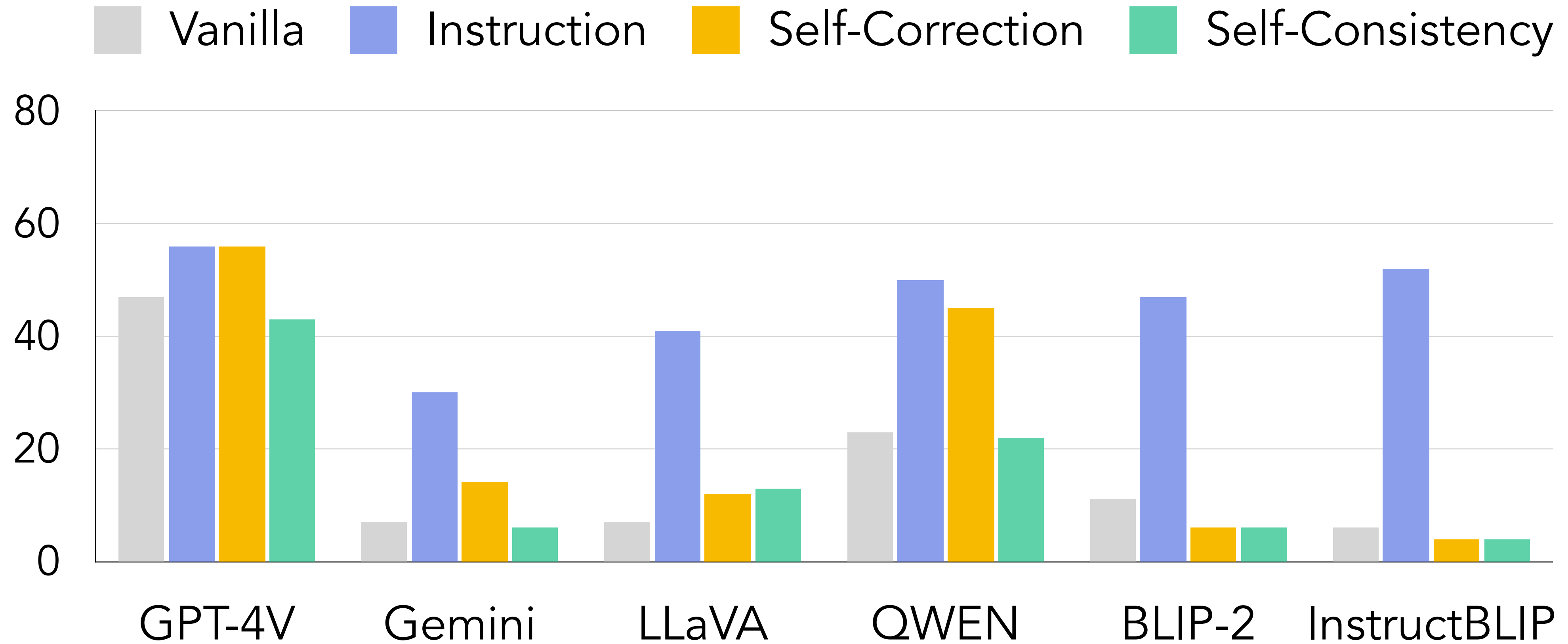
How can we automatically evaluate long-form answers?

Pearson Correlation with Human ratings



How can models abstain to unanswerable questions?

Abstention (%)



How can models abstain to unanswerable questions?

Abstention (Recall)

