ARITHMETIC REASONING
Complementary Multimodal Chain-Of-Thought

10 Cross-referencing Categories

OCR | External Knowledge | Distractors | Hallucination | ...

Example

Eliza's pay for the first 40 hours she works each week is the same value of the money bills in the picture. She also receives an overtime pay of 1.2 times her regular hourly rate. If Eliza worked for 45 hours this week, how much are her earnings for this week?

2600 image-question pairs

GPT-4o: 62.18% (accuracy)

MULTI-IMAGE REASONING
Multi-image VQA

3 Categories

Arithmetic Reasoning | Cause & Effect | External Knowledge

Example


Kylar went to the store to buy glasses for his new apartment. One glass costs $5, but every second glass costs only 60% of the price. Kylar wants to buy the number of glass held by the heroes plus twice the number of glasses held by the villains as in the pictures. How much does he need to pay for them?

316 image-question pairs

GPT-4o: 57.89% (accuracy)

CAPTIONING
Imaginary Image Captioning

11 Categories

Unexpected Behavior | Misplacement | Fictional Environment | ...

Example

1. The image depicts an elephant with a unique building entry and a wall clock incorporated on its body, situated in the middle of a desert. **2.** An elephant is seen bearing elements of architecture, with a door and a clock tower on its body, in a desert backdrop. **...** **5.** The picture is of an elephant showing a case of surrealism featuring a built-in clock tower and door, standing alone amidst the sand dunes.

1000 images – 5000 captions

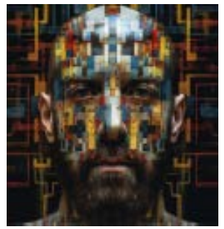
GPT-4o: 32.56 (CIDEr)

VISUAL QUESTION ANSWERING
VQA with Hallucination Triggers

3 Categories

Insufficient Context | False Premise | Visually Challenging Images

Example



False Premise Prompt

Are the man's earring made out of gold or silver?

Insufficient Context

What color are the shoes worn by the woman in the red dress?

Visually Challenging Images

Is the man sitting on a stool or a chair?

7748 image – question pairs

GPT-4o: 68.10% (accuracy)

CROSS-MODAL RETRIEVAL
Fine-grained Retrieval with Sample-specific Distractors

11 Categories

Example

I2T with Sample-specific distractors

The cats are standing on their hind legs and appear to be dancing outdoors.

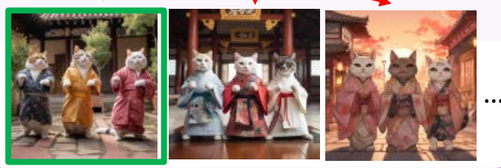
The cats are standing on four of their legs and dancing.

The cats are all wearing traditional kimonos of the same color.

...

T2I with sample-specific distractors

The cats are standing on their hind legs and appear to be dancing outdoors.



Text retrieval: 11121 text, 1000 images

Image retrieval: 5000 text, 6323 images