

NL-Eye

Abductive NLI For Images

ICLR 2025



Mor Ventura, Michael Toker, Nitay Calderon, Zorik Gekhman,
Yonatan Bitton, and Roi Reichart

Will a VLM-based bot warn us about slipping if it detects a wet floor?

Premise



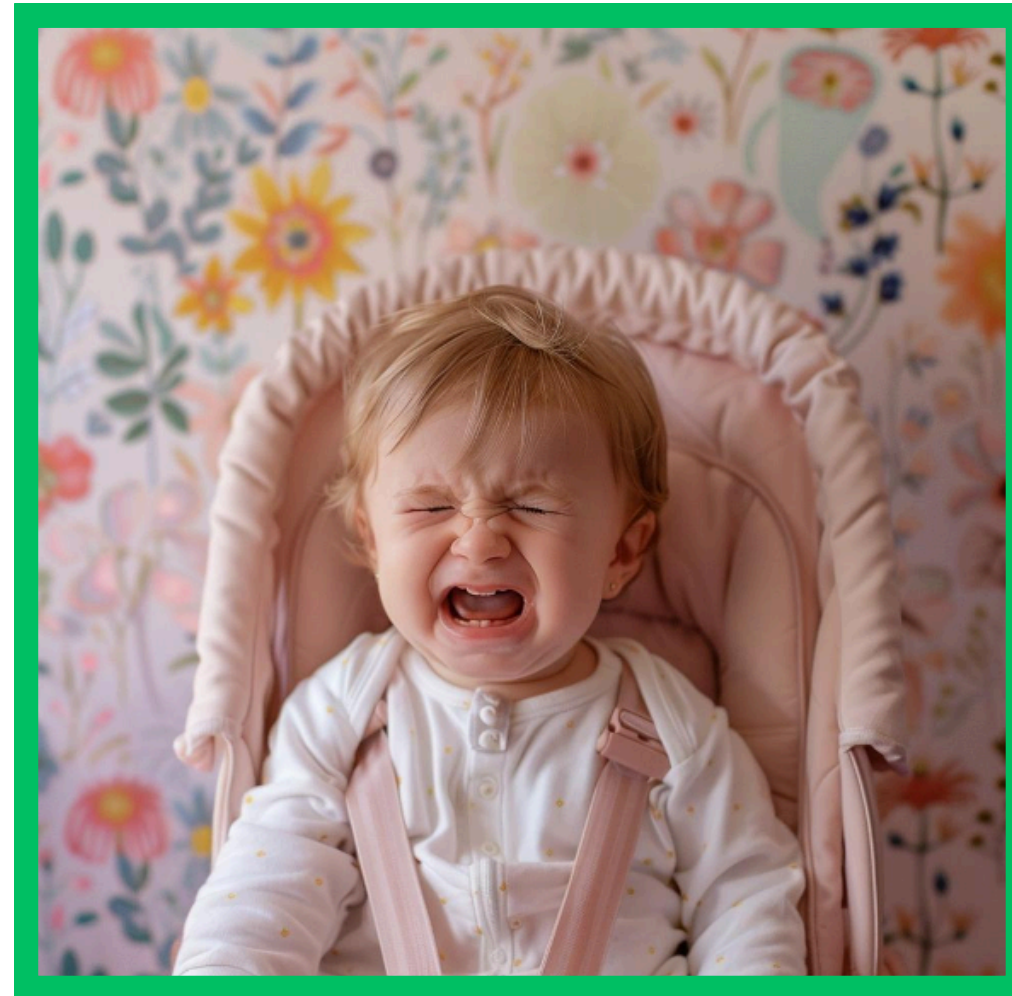
Plausible Hypothesis



Implausible Hypothesis





Would it infer a missing pacifier as a cause of a crying baby?



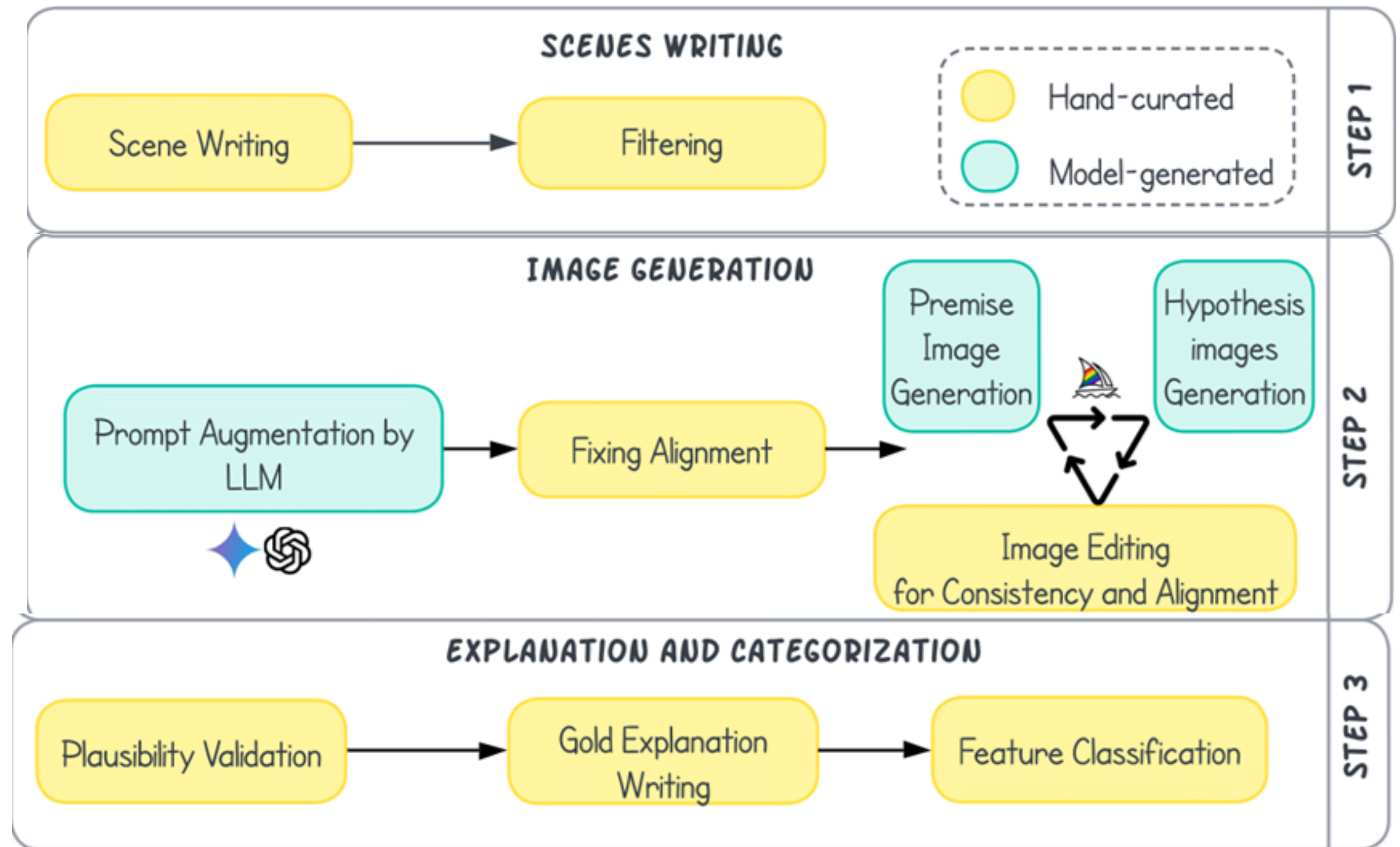
*Temporal direction: forward

Existing Sources

Parameter \ Source	Multiple Images	Scenes Flexibility	Abductive Reasoning
 Videos	✓	✗	✗✓
Textual NLI		✗	✗✓
 Visual Reasoning	✗	✗	✗✓
NL-Eye	✓	✓	✓

Bitton-Guetta et al, 2023, Breaking Common Sense: WHOOPS!
Nie et al, 2019: Adversarial NLI: A New Benchmark for Natural Language Understanding
Li et al, Nov 2023, SEED BENCH 2

NL-Eye: Dataset Curation Workflow



NL-Eye Tasks

1. Plausibility Prediction

Which image is **more plausible**?

(2) Plausibility Explanation

And **why**?

Prompts

Premise

A BASKETBALL TEAM IN
PURPLE UNIFORMS WINS
THE GAME.

Plausible Hypothesis

A CHILD WITH A PURPLE
SCARF CHEERS.

Implausible Hypothesis

A CHILD WITH A PURPLE
SCARF WORRIES.

Text-to-Image



STEP 1

STEP 2

Explanation

WHY IS THIS SCENE MORE PLAUSIBLE?

The child is likely to be happy because they cheered for the purple team.

STEP 3

Features



Emotional

Reasoning Category



Short term

Temporal Duration



Forward

Temporal Direction

NL-Eye Benchmark: 350 triplets as test set

PHYSICAL

Premise



Plausible Hypothesis



Implausible Hypothesis



Because banana rot after a month

FUNCTIONAL

Premise



Plausible Hypothesis



Implausible Hypothesis



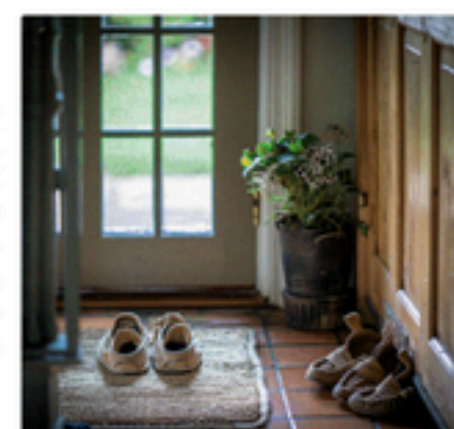
Because the rolling pin flattens the dough

LOGICAL



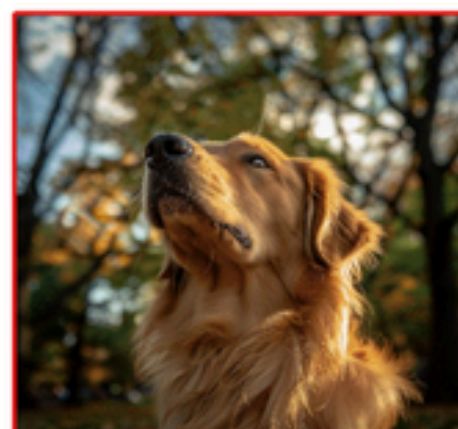
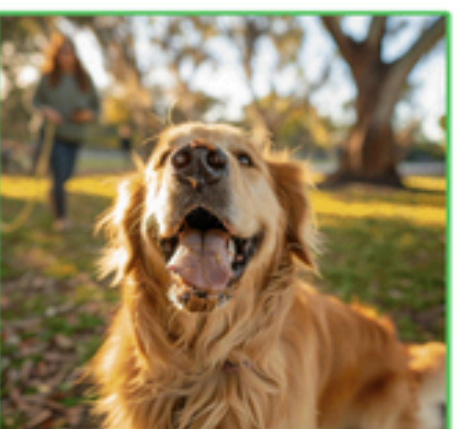
Homes with children typically have a fridge stocked with kid-friendly food

CULTURAL



Wearing house shoes is a common Russian trait, similar to having Matryoshka dolls

EMOTIONAL



Treats are positive feedback that makes dogs happy

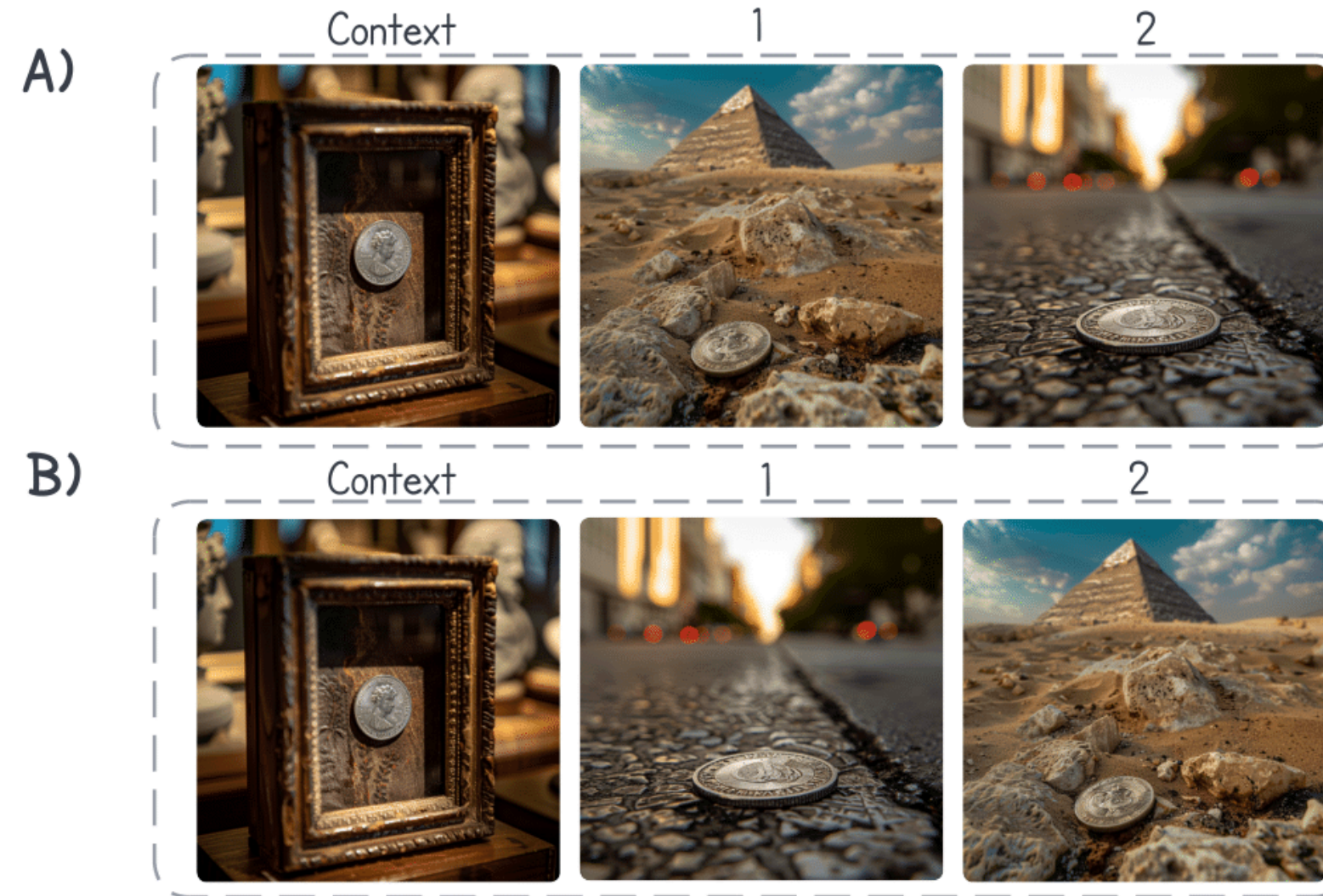
SOCIAL



Because a family photo is more likely to indicate a married person

Setup

Triplet Setup



A) More Plausible: **Image 1**

Explanation: **Museum coins typically have significant archaeological origins**

&

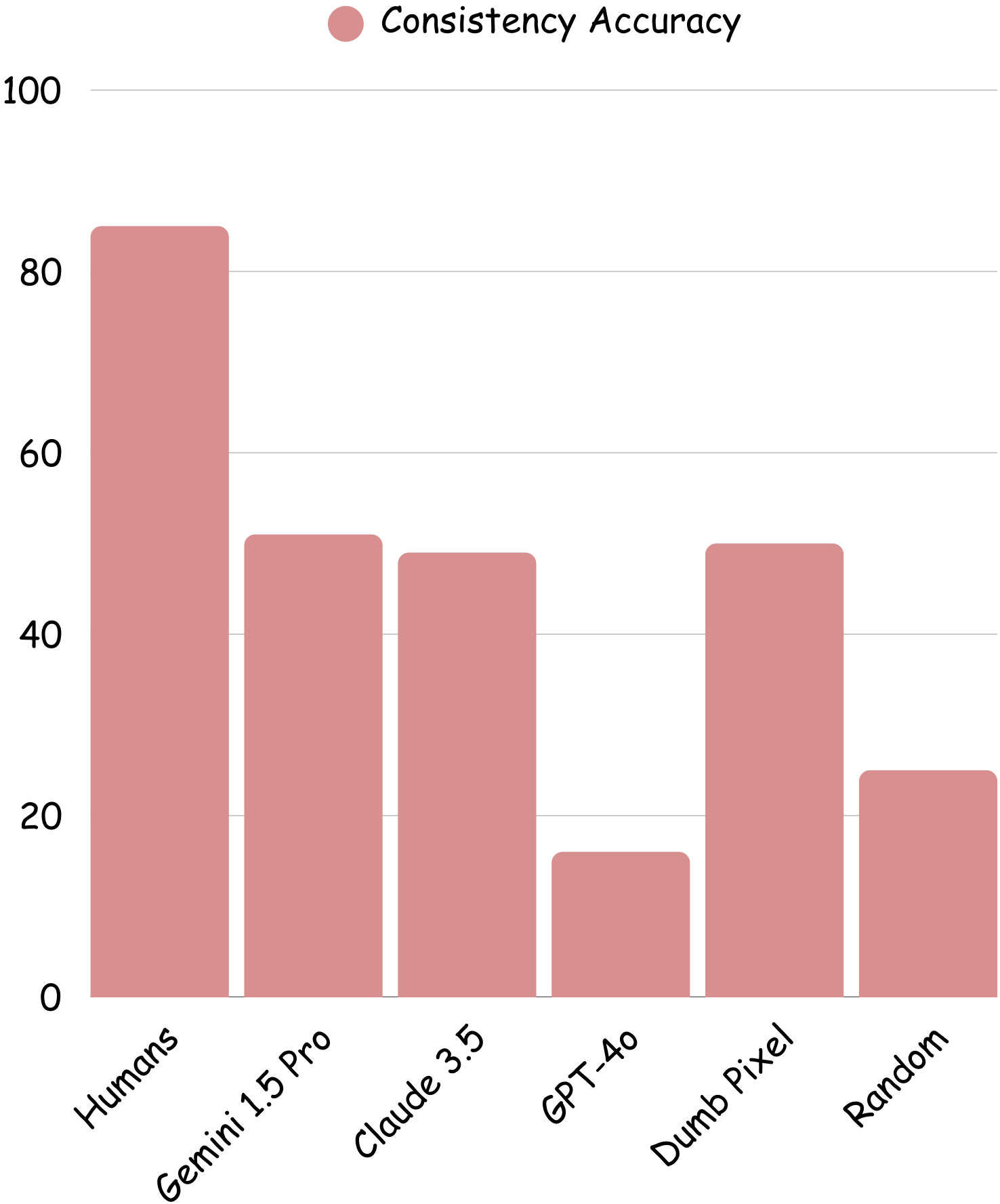
B) More Plausible: **Image 2**

Explanation: **Findings from the Pyramids archaeological site are displayed in museums**



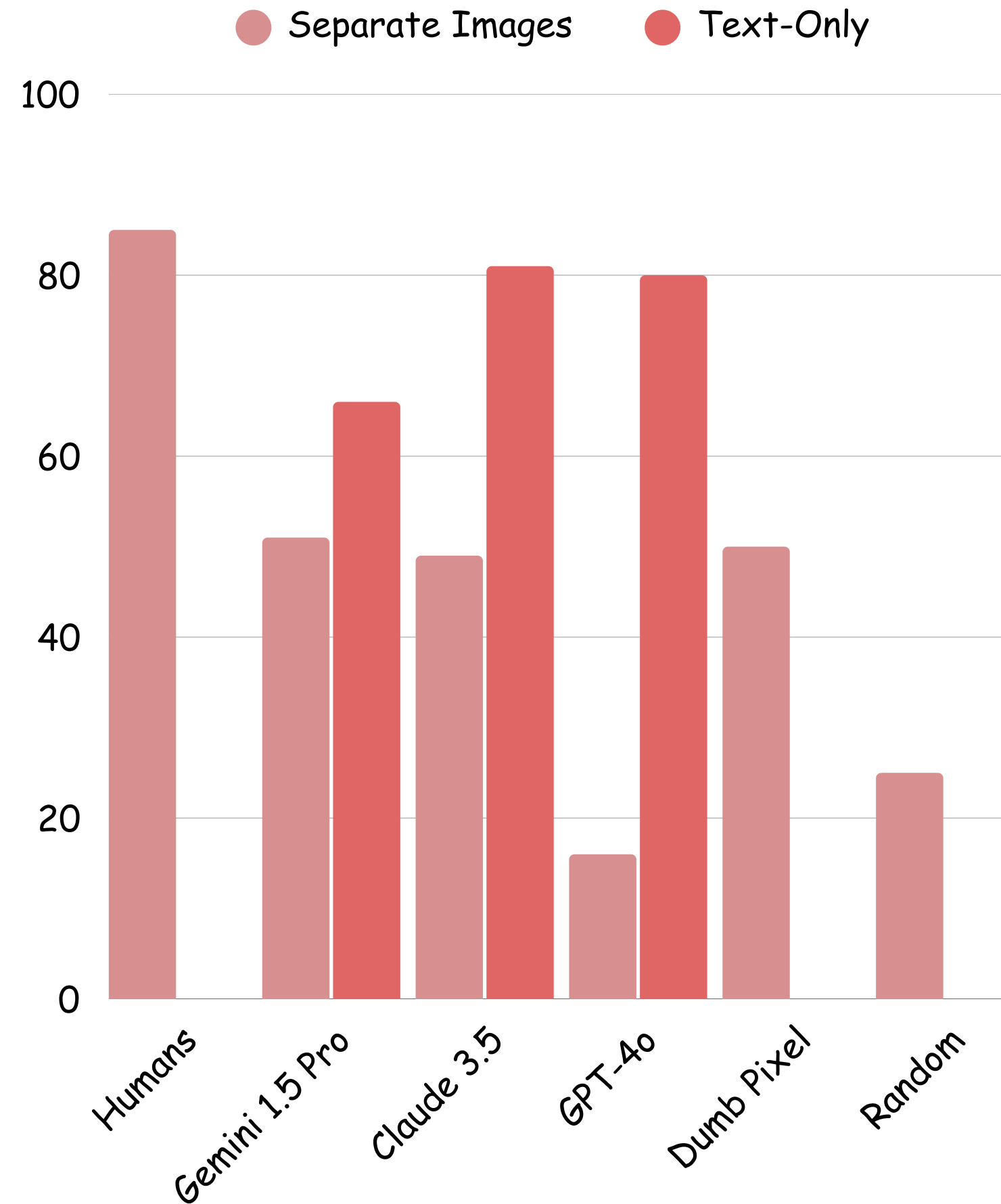
Consistency-Accuracy

(1) VLMs Fail Where Humans Excel

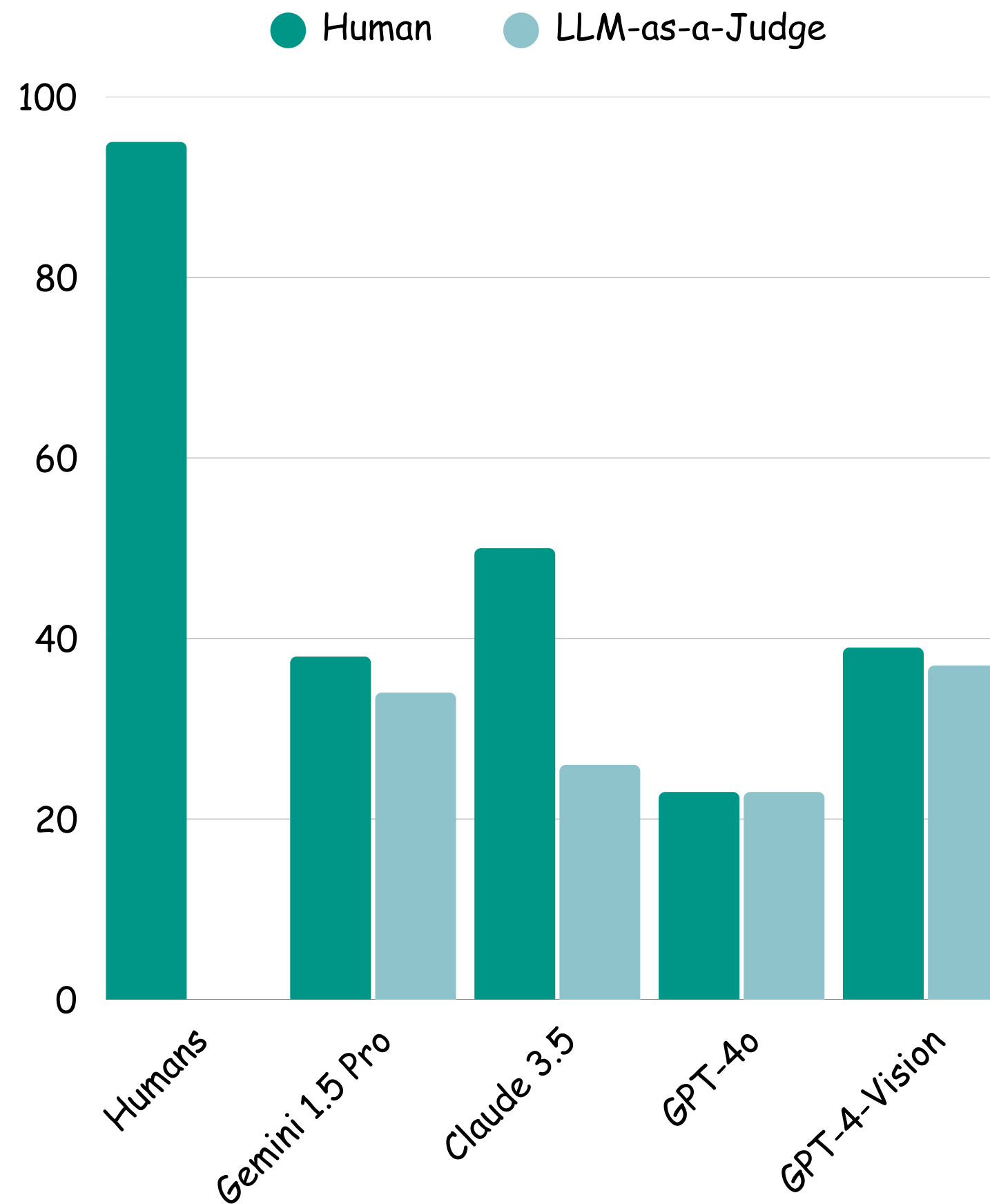


*greedy decoding results

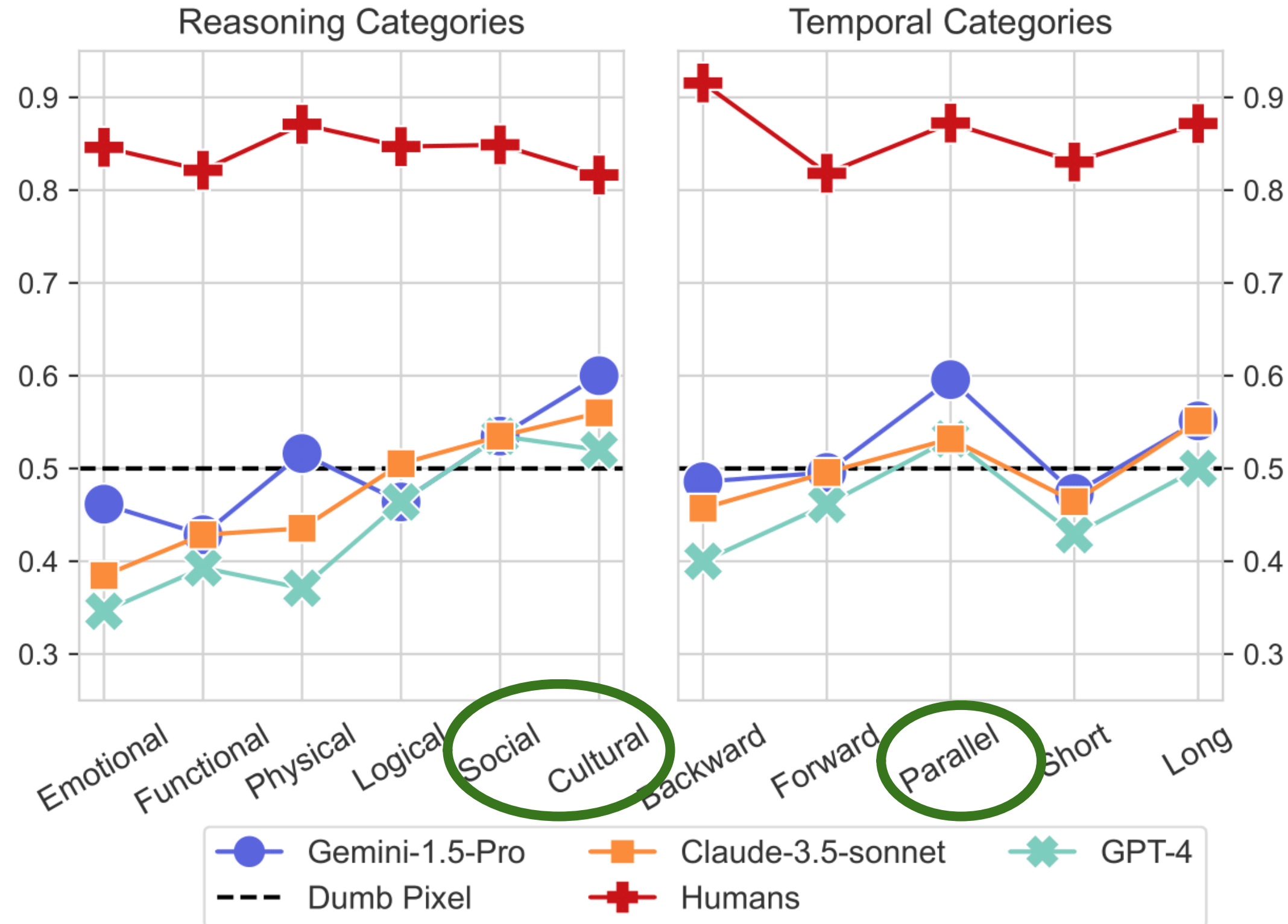
(2) The Failure is in the Visual Interpretation



(3) Even When Correct, VLM Explanations Are Unhelpful



(4) VLMs are Better in **Correlational** and **Knowledge-based** Reasoning Compared to Causal Reasoning



NL-Eye: Abductive NLI For Images

- (1) VLMs Fail Where Humans Excel**
- (2) The Failure is in the Visual Interpretation**
- (3) Even When Correct, VLM Explanations Are Unhelpful**
- (4) VLMs are Better in Correlational and Knowledge-based Reasoning**
- (5) Failure Factors in VLMs Explanation**

