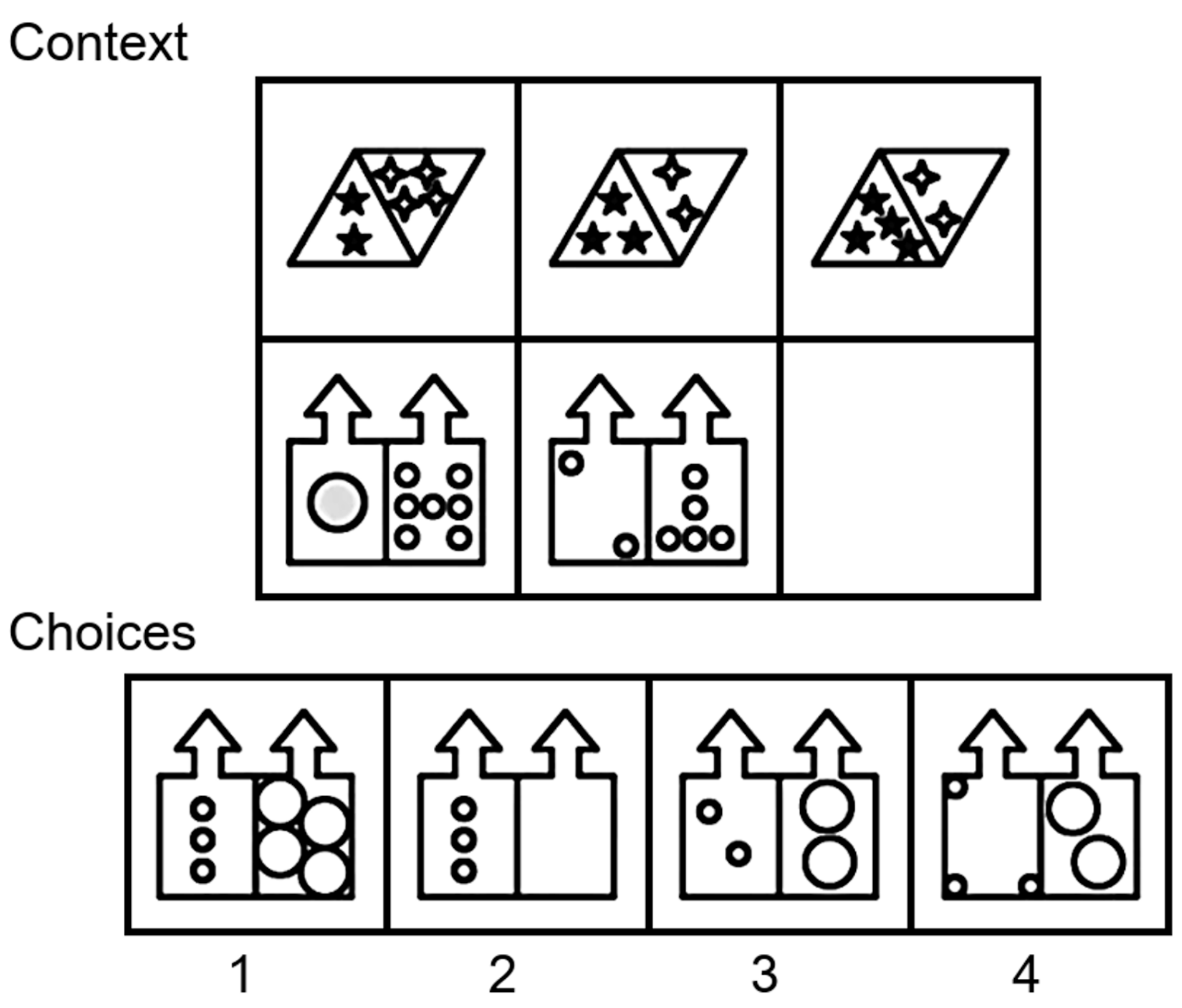
# MARVEL: Multidimensional Abstraction and Reasoning through Visual Evaluation and Learning

Yifan Jiang[1*], Jiarui Zhang[1*], Kexuan Sun[1*], Zhivar Sourati[1],
Kian Ahrabian[1], Kaixin Ma[3], Filip Ilievski[2], Jay Pujara[1]

[1]Information Sciences Institute, University of Southern California
[2]Vrije University Amsterdam
[3]Tencent AI Lab
*Equal Contribution

Github Paper
Website Huggingface

# Pause and ponder: A puzzle no MLLM has solved yet!



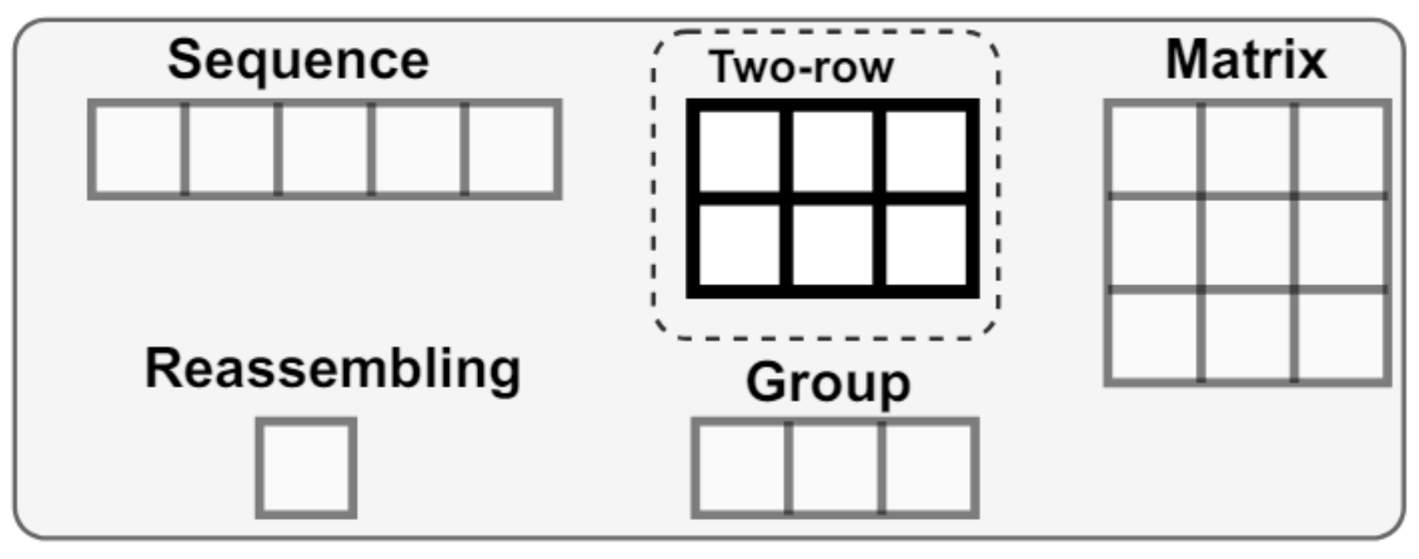
## Research Goal:

- Marvel: **A multi-dimensional abstract visual reasoning benchmark** evaluating MLLMs across different patterns, input shapes, and task configurations.
- **A hierarchical evaluation framework** incorporating perception questions with AVR questions to enable fine-grained diagnosis of model capability.

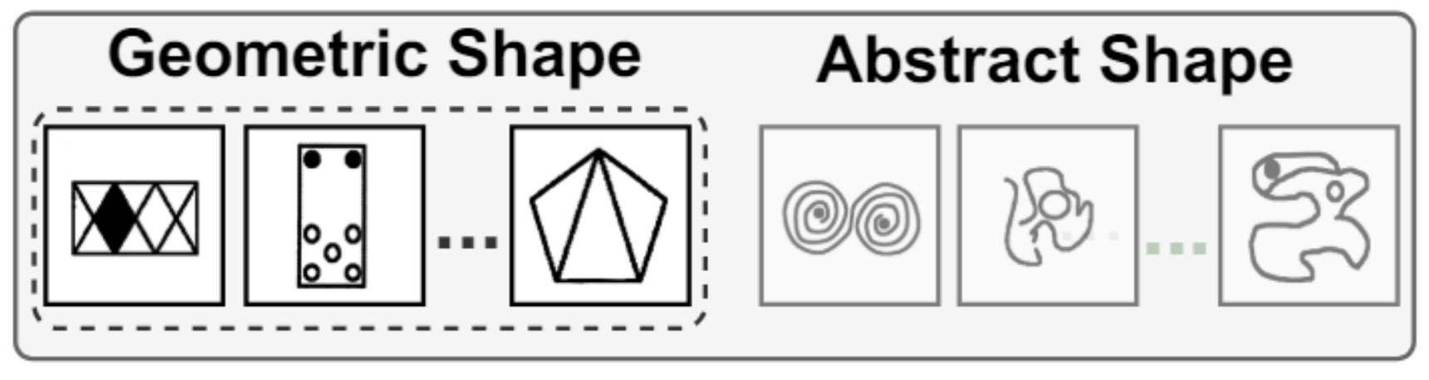
## MARVEL Benchmark Construction

Each puzzle in MARVEL consists of:

•**Task Configuration** arranges the panels.



•**Input Shapes** in each panel.



•**Pattern** governs the relationship between puzzle panels.

Expand from core knowledge

◆ Object Core Knowledge: 1)Temporal Movement; 2)Spatial Relationship

◆ Number Core Knowledge:1)Quantities; 2)Mathematical

◆ Geometry Core Knowledge: 1)2D-Geometry; 2)3D-Geometry

## Hierarchical Evaluation Framework

**AVR Question**: The puzzle consists of a question part in a 2 by 3 matrix with the right piece missing. Which choice is the answer to fill the missing part?

↓

Coarse-grained Perception Question: How many panels, including blank panel, in the Context part?

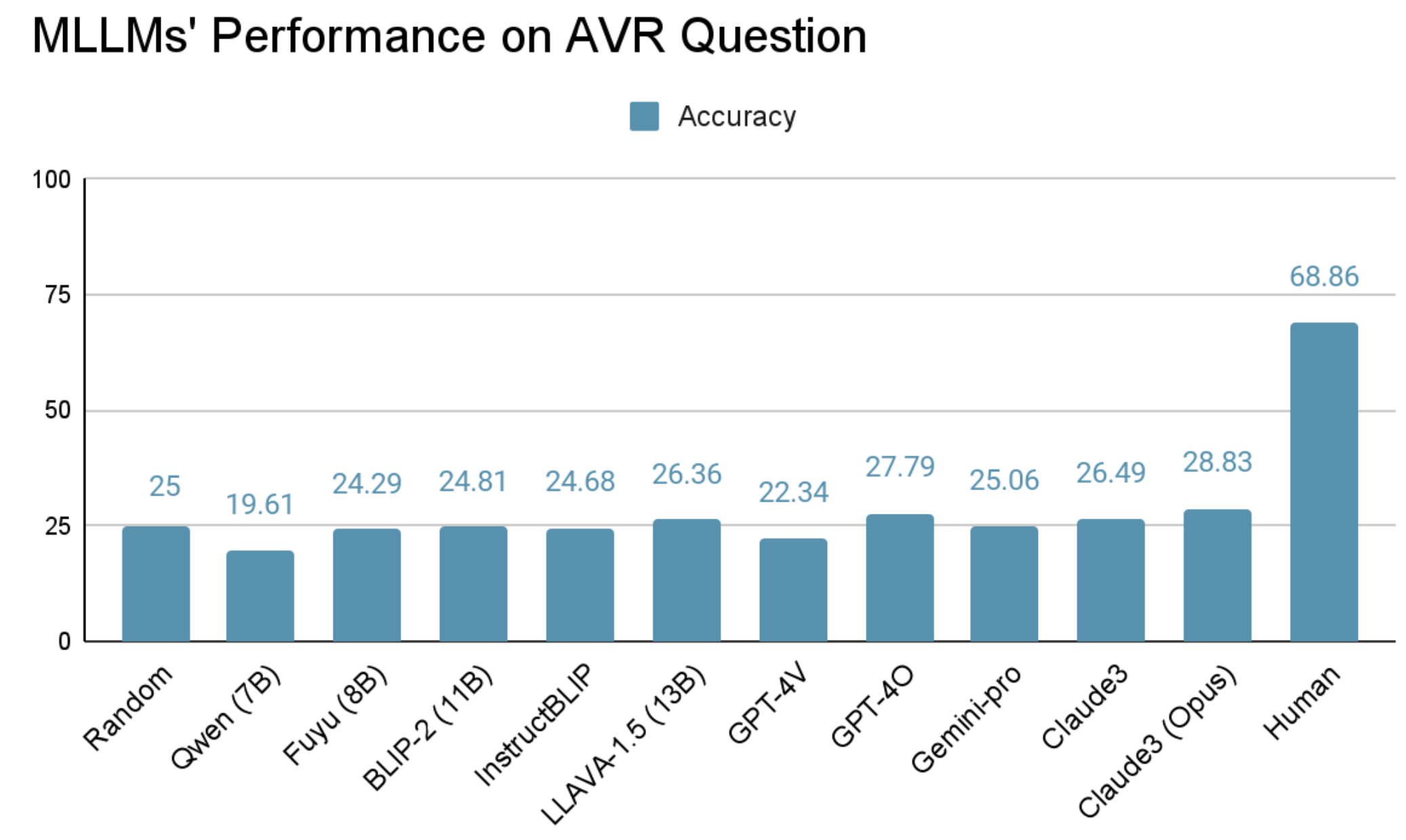
1) Context part 2) Choice part 3) Whole puzzle

↓

Fine-grained Perception Question: Are there two or three circles in the left half of choice 4?

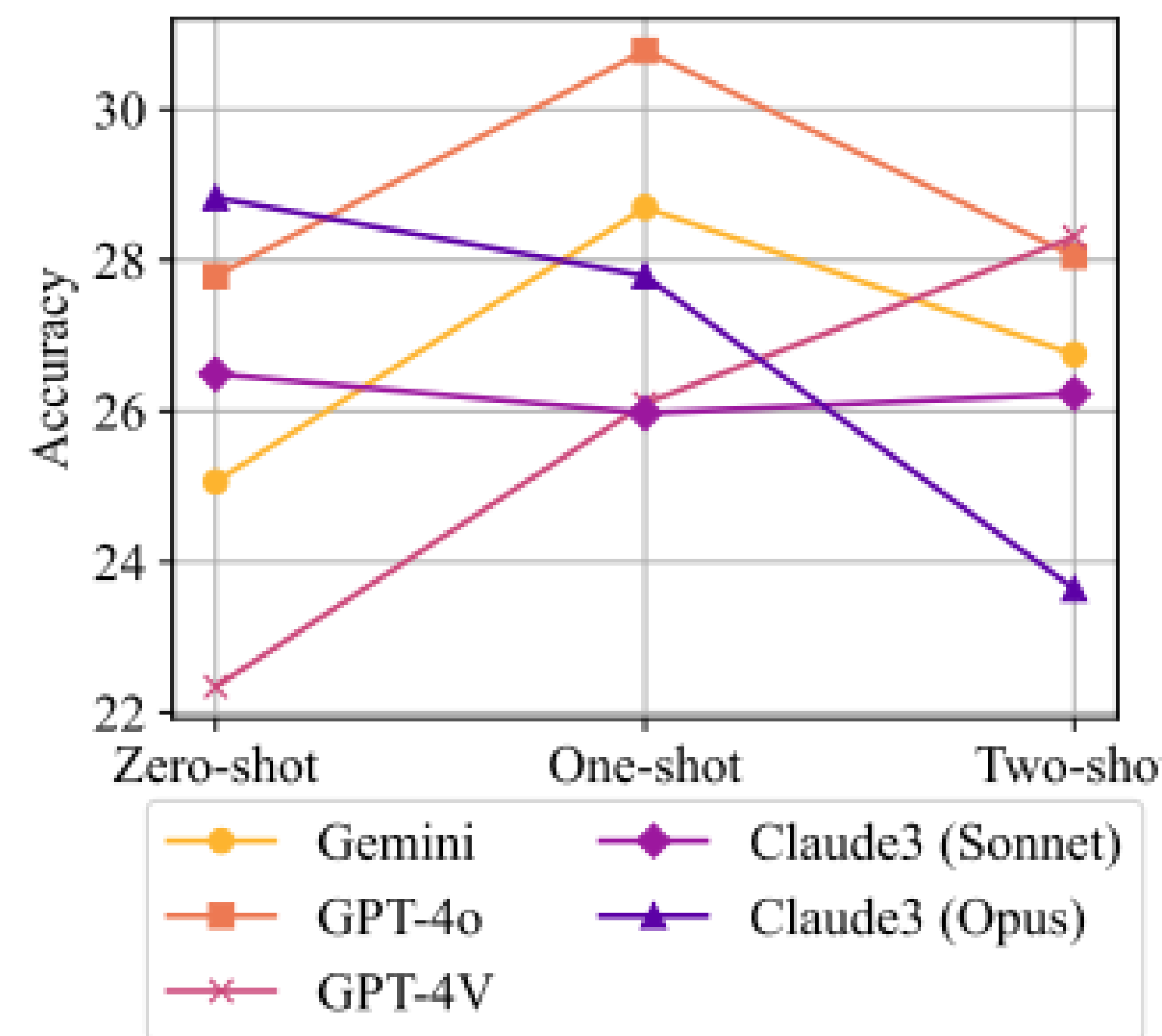
1) Location 2) Color 3) Shape 4) Quantity 5)Comparison

## What's the abstract reasoning ability on visual puzzles of current SOTA MLLMs?



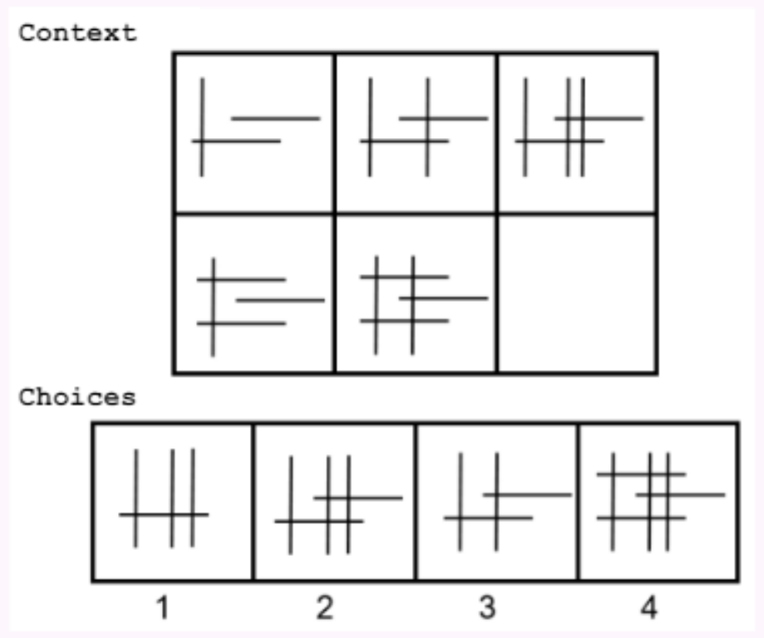
- All models show near-random performance with a huge gap (40%) compared to human performance
- Extremely imbalanced distribution in the outputs of some MLLMs

## Can MLLMs do better with different few-shot prompting strategies?



- The few-shot demonstrations show a marginal positive impact on GPT-4V and a decreasing trend on Claude3

## Perception question, zero- and two-shot example of Claude3 (Opus)



**Perception Question:**
Are there two or three parallel horizontal lines in choice 2?

**Claude3-Opus:**
Choice 2 contains three parallel horizontal lines.

**Claude3-Opus (Zero-shot):**
To determine which choice best fits the missing piece in the bottom right, we need to identify the pattern in the given 3x2 grid. Looking at the grid, we see that each row follows a specific pattern:

Row 1: The number of vertical lines increases by 1 from left to right.
Row 2: The number of horizontal lines increases by 1 from left to right.
Based on this pattern, the missing piece in the bottom right should have 3 horizontal lines, as it follows the pattern in the second row.

Looking at the choices provided, choice 3 is the only option that has 3 horizontal lines. Therefore, **choice 3** is the most appropriate answer to fill the missing part in the given puzzle.

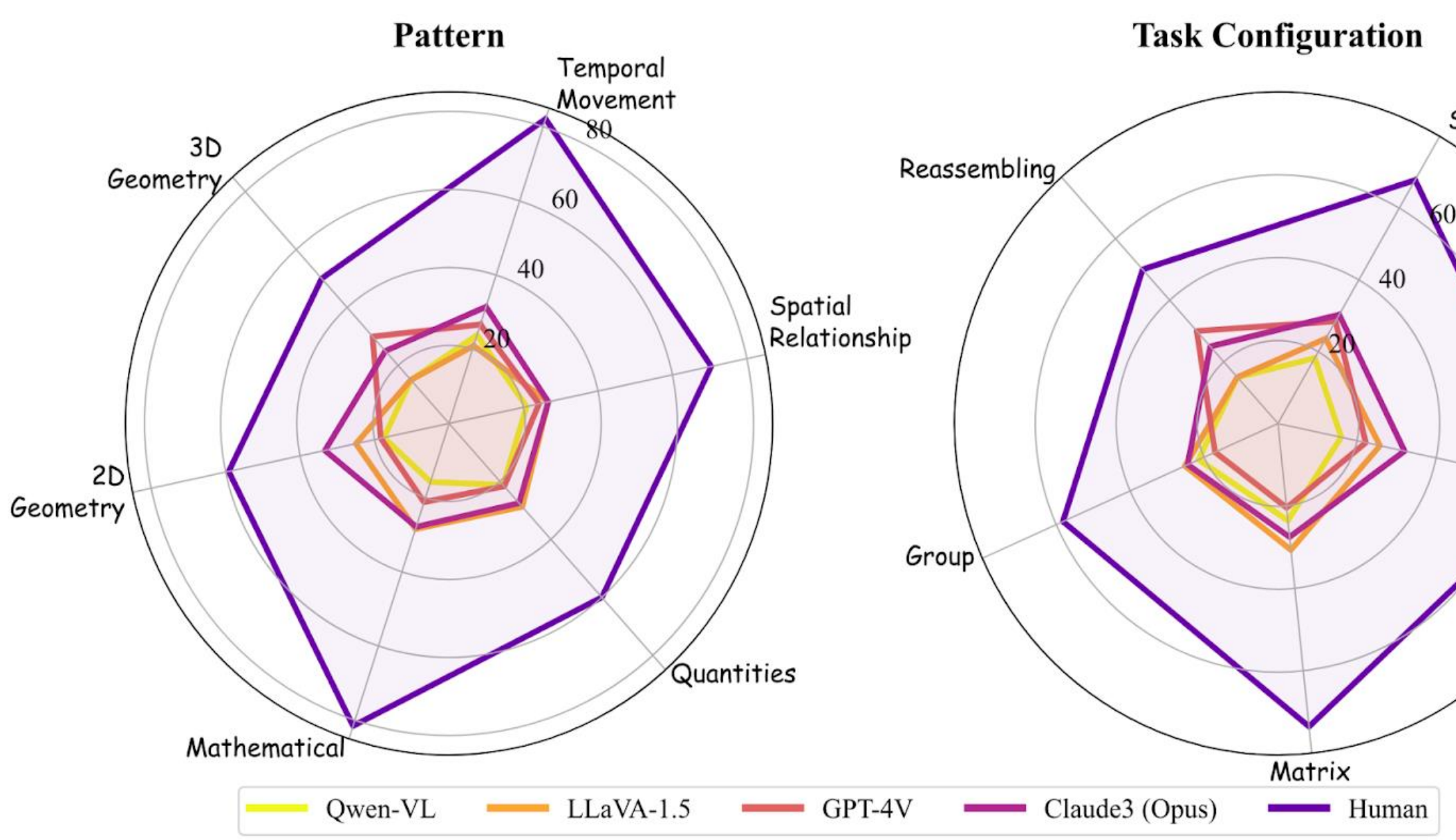
**Claude3-Opus (Two-shot):**
Let's analyze the pattern in the given 3x2 matrix, focusing on the number of vertical and horizontal lines in each cell

Row 1: Cell 1 has one vertical line, Cell 2 has two vertical lines
Row 2: Cell 1 has one vertical and one horizontal line, forming a plus sign. Cell 2 has two vertical and two horizontal lines, forming a hash symbol.
Row 3: Cell 1 has three vertical lines. The missing cell should follow the pattern established in the previous rows.

Based on the pattern, the missing cell should have three vertical and three horizontal lines, forming a grid of nine squares. This corresponds to choice 4 among the given options. Therefore, **choice 4** is the most appropriate answer to complete the matrix.

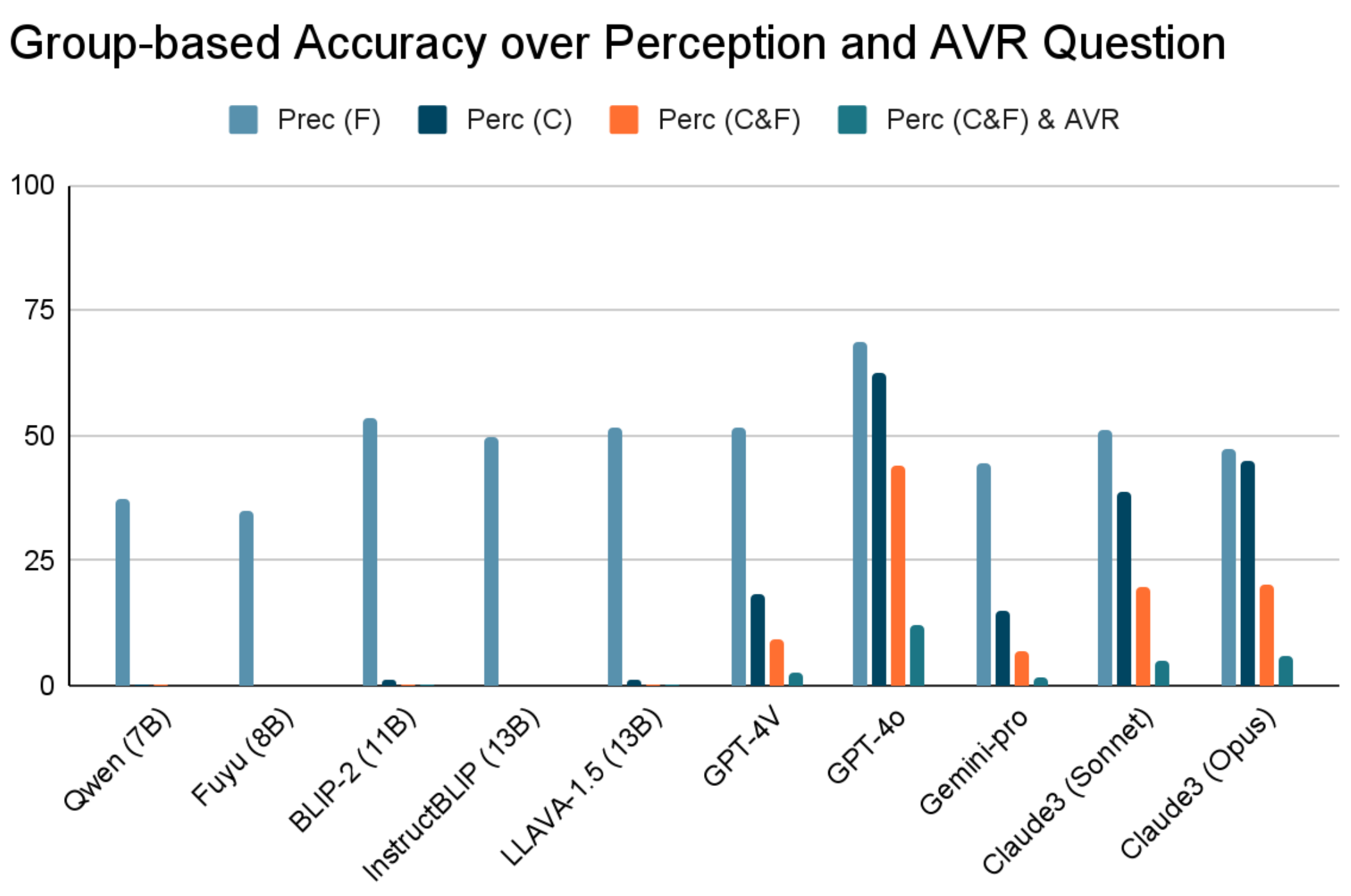
- The model's reasoning is based on the perception of the puzzle, which needs to be completely precise to support correct reasoning.
- A single error in visual feature perception can impact reasoning since the correct pattern must apply to all puzzle shapes.

## How do MLLMs perform on different patterns and task configurations?



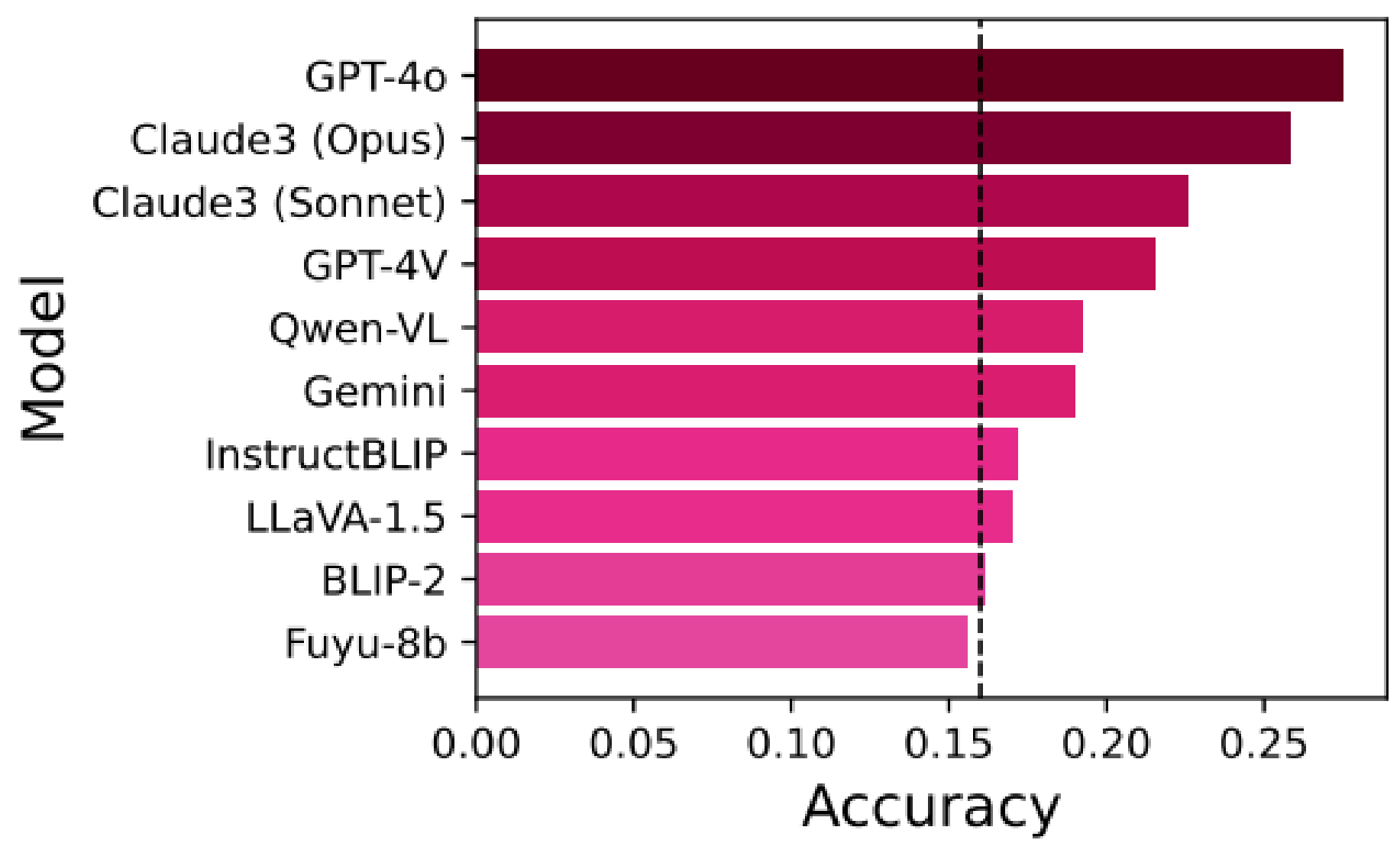
- 3D-Geomertry pattern is the most challenging
- 2D-Geomertry seems relatively easier
- Four out of five MLLMs rank 1st in different task configurations

## Do MLLMs visually understand puzzle, and do they show consistent reasoning ability?



- All models show near-random performance on visual detail perception，
- No model can solve the AVR puzzles with consistent reasoning.

## Models' potential when perceptual barriers are mitigated



Asking for possible underlying patterns in a multiple-choice

Introducing text description in the input

| Input | Open-sourced MLLMs | | | | | Closed-sourced MLLMs | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Qwen-VL | Fuyu | Blip-2 | InstructBLIP | LLaVA-1.5 | GPT-4V | GPT-4o | Gemini | Claude3 (Sonnet) | Claude3 (Opus) |
| AVR | 23.16 | 23.16† | 23.16† | 23.16† | 21.05 | 21.05 | 23.16 | 26.32† | 27.37† | 30.53 |
| AVR+Text | 24.21↑1.05 | 26.32↑3.16 | 26.32↑3.16 | 13.68↓9.48 | 28.42↑7.37 | 65.26↑44.21 | 58.95↑35.79 | 37.89↑11.57 | 49.47↑22.10 | 55.79↑25.26 |