Vision Language Models are Blind

AUBURN UNIVERSITY · UNIVERSITY OF ALBERTA

Pooyan Rahmanzadehgervi*, Logan Bolton*, Mohammad Reza Taesiri, Anh Totti Nguyen

Project , code, data

vlmsareblind.github.io

Summary

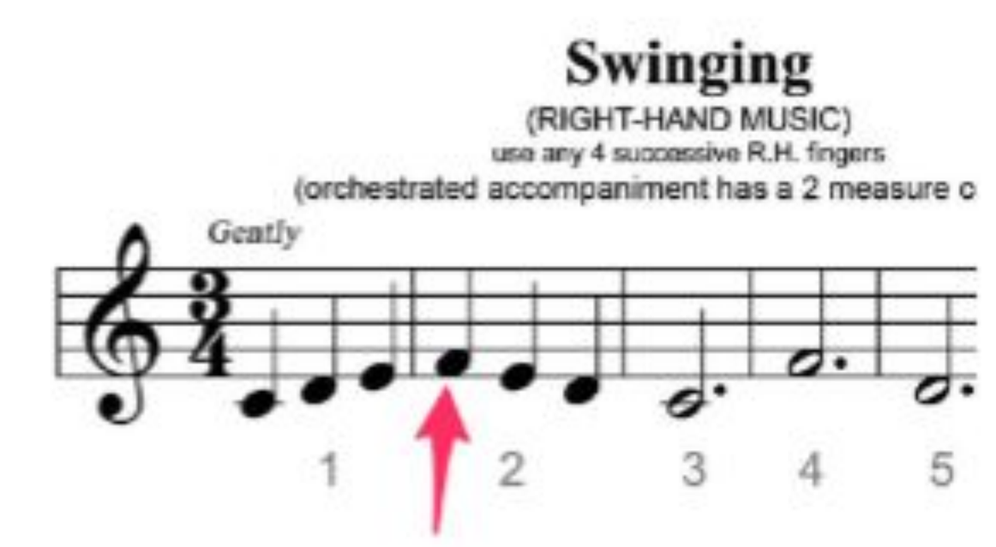
Large language models with vision capabilities *still* **struggle** with **low-level vision tasks** that are trivial to humans.

Motivation

- Gemini can **solve 42.9%** of the questions in MMMU benchmark **without seeing the input image**. Are we on the right way…Chen et al. 2024
- Most VQA benchmarks **do not exclusively** test **vision** capabilities.
- Text-only LLMs can reach **> 80%** of SOTA on DocVQA, TextVQA, ChartQA, AI2D (images are serialized). Analyzing…Hedge et al. 2023

How would VLMs perform on tasks that exclusively test vision?

Reading music sheets



User: What is the name of this note?
Sonnet-3.5: The note indicated by the pink arrow is F. ✓ This can be determined by its position on the musical staff - it sits on the top line of the treble clef, which corresponds to the note F.

User: Is the note on a line?
Sonnet-3.5: Yes, this note is on a line. ✗ Specifically, it's on the third line from the bottom of the treble clef staff. ✗

Reading NYC street map

User: Is Catherine St intersecting with Market St?

GPT-4o: Yes, Catherine St does intersect with Market St as shown in the map ✗. They intersect near the Alfred E. Smith Playground and close to the Monroe St and Cherry St intersections. ✗

Gemini-1.5: No, Catherine St and Market St do not intersect in this map. ✓

Findings

VLMs **cannot** reliably:

1. Tell if two circles are touching
2. Count the number of times two lines intersect
3. Follow paths from one point to another
4. Reliably count how many rows and columns are in a table
5. Tell which letter is being circled

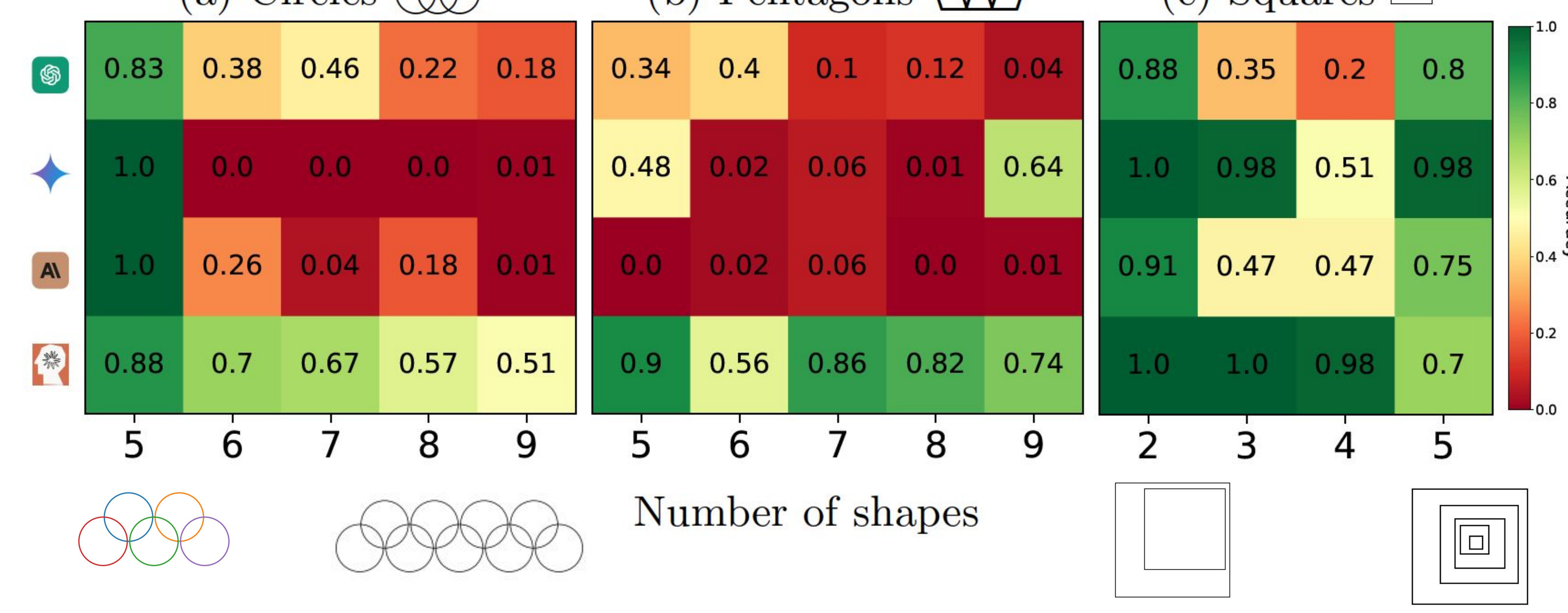
58% accuracy

VLMs **can accurately perform** these same tasks when the shape primitives are distanced away.

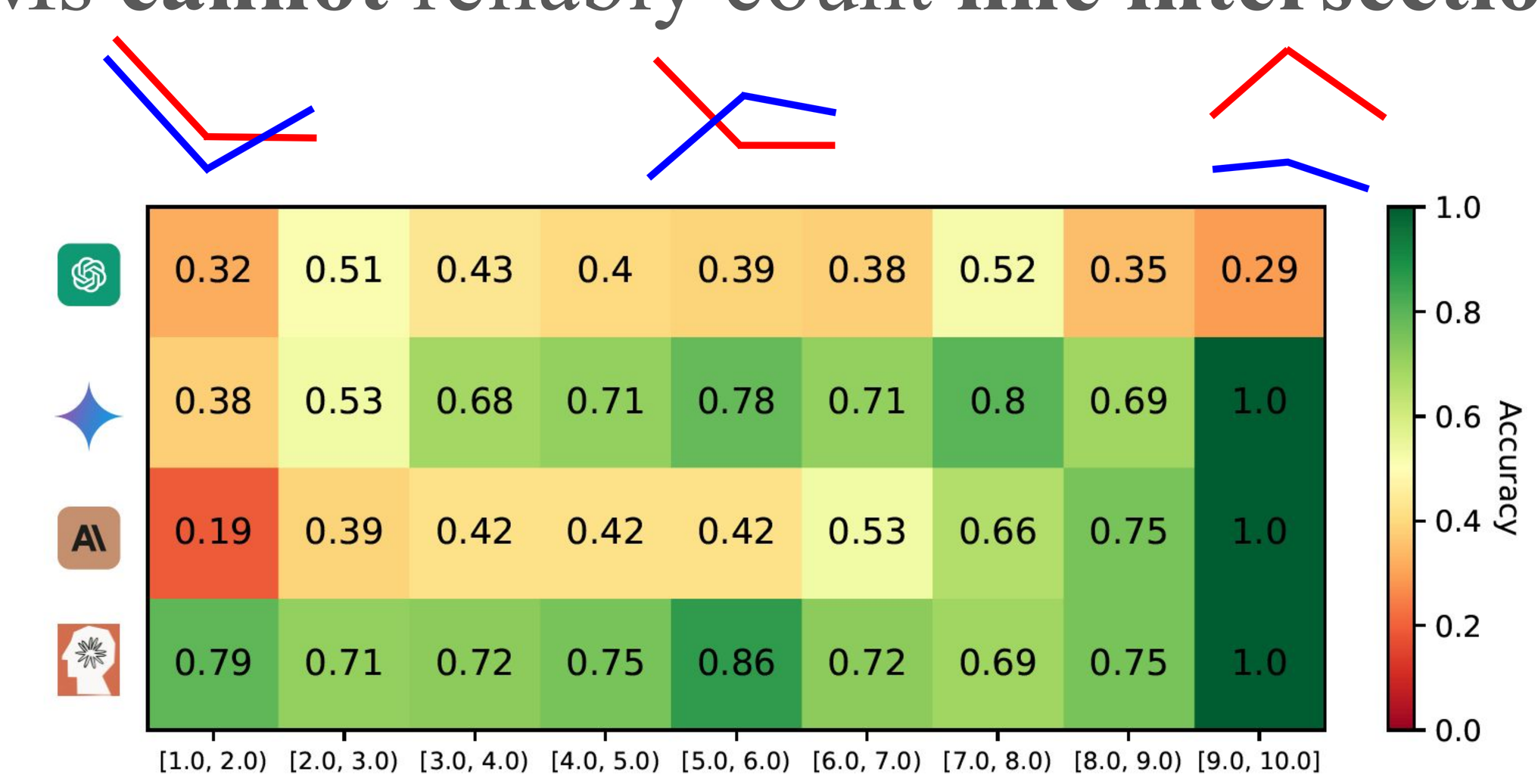
Model									Task mean
Random	33.33	50.00	5.77	20.00	20.00	25.00	4.55	33.33	24.00
GPT-4o	41.61	75.91	74.23	41.25	20.21	55.83	39.58	53.19	50.23
Gemini 1.5 Pro	66.94	**93.62**	83.29	20.25	24.17	87.08	39.39	53.13	58.48
Sonnet 3	43.41	86.46	72.06	29.79	1.87	65.00	36.17	31.11	45.73
Sonnet 3.5	**75.36**	90.82	**87.88**	**66.46**	**77.71**	**92.08**	**74.26**	**58.19**	**77.84**
Mean	56.83	86.70	79.37	39.44	30.99	75.00	47.35	48.91	58.07

Results

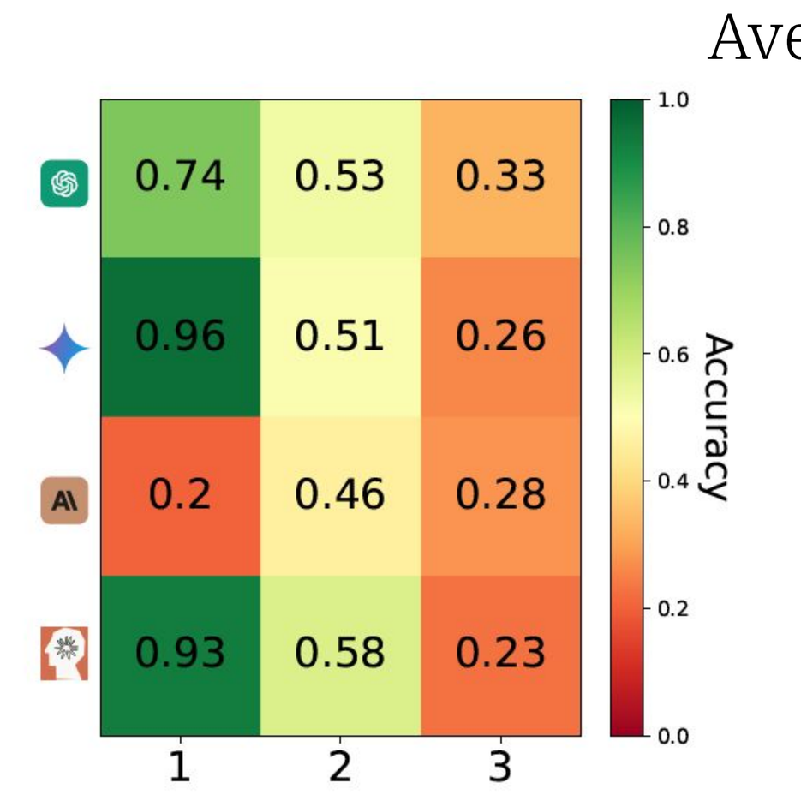
VLMs **struggle** to **count overlapped** and **nested** shapes

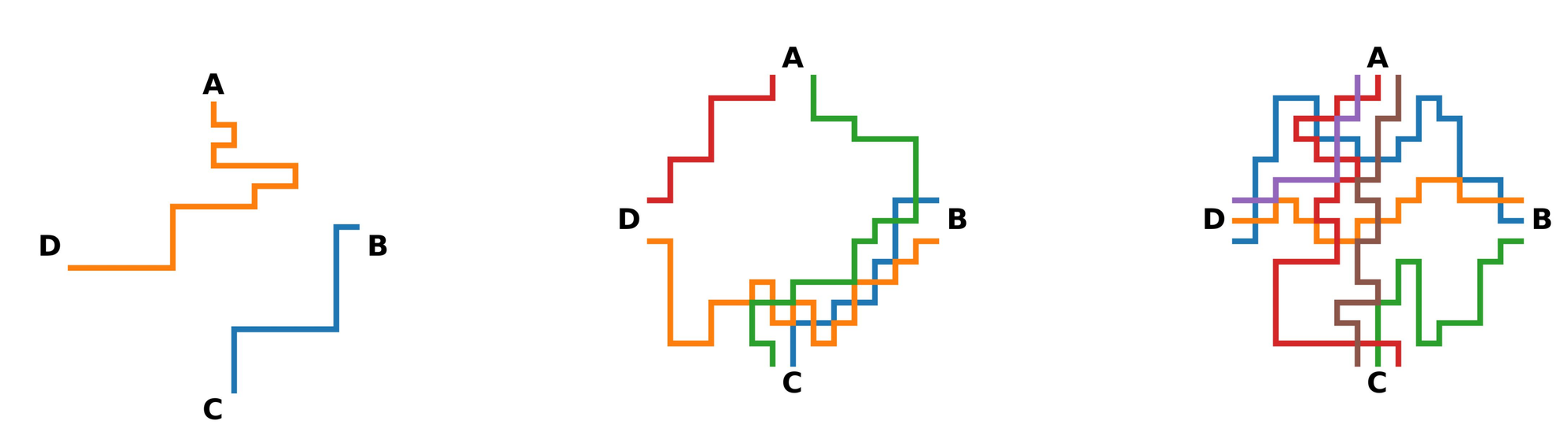


VLMs **cannot** reliably count **line intersections**



VLMs **struggle** to **count single-colored** paths





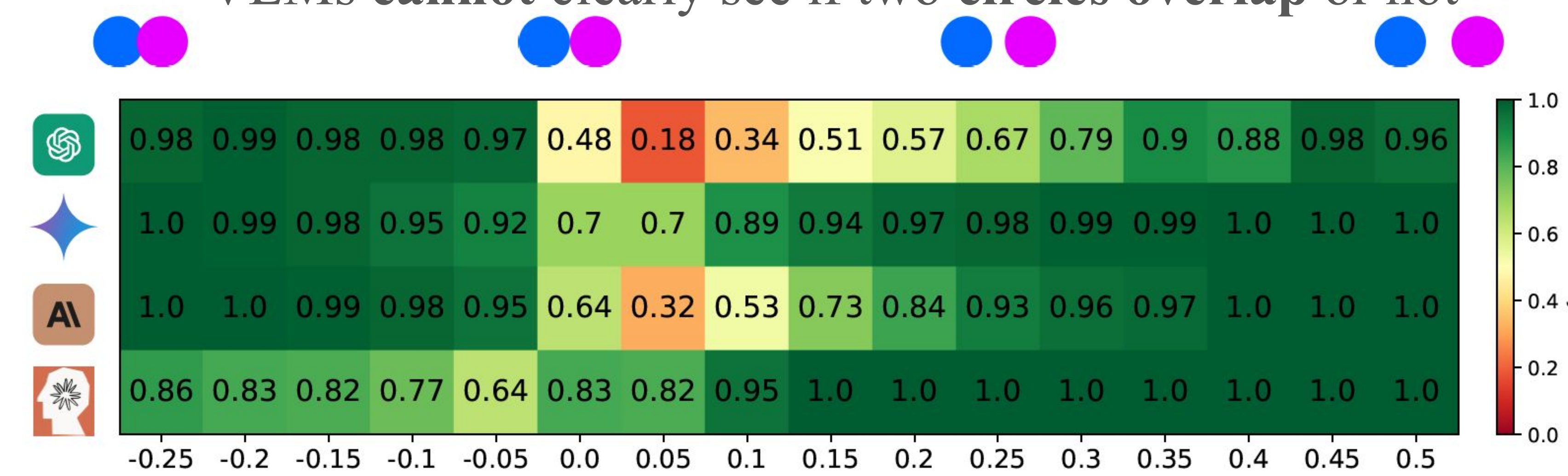
Examples from BlindTest benchmark with VLMs' responses

	P1		P2		P3		P4		P5		P6		P7	
GPT-4o	1	✗	Yes	✗	o	✗	6	✓	5	✗	3×4	✓	1	✓
Gemini-1.5	1	✗	No	✓	w	✗	5	✗	3	✓	3×4	✓	2	✗
Sonnet-3	1	✗	Yes	✗	o	✗	5	✗	4	✗	4×4	✗	2	✗
Sonnet-3.5	0	✓	No	✓	1	✓	6	✓	3	✓	3×4	✓	1	✓

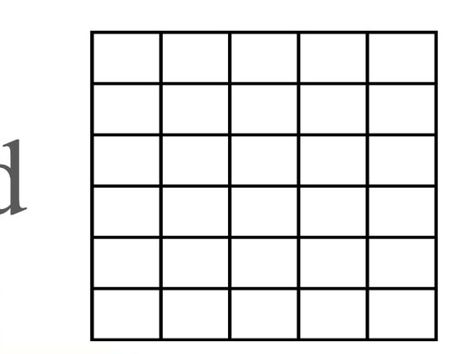
GPT-4o · Gemini-1.5 · Sonnet-3 · Sonnet-3.5

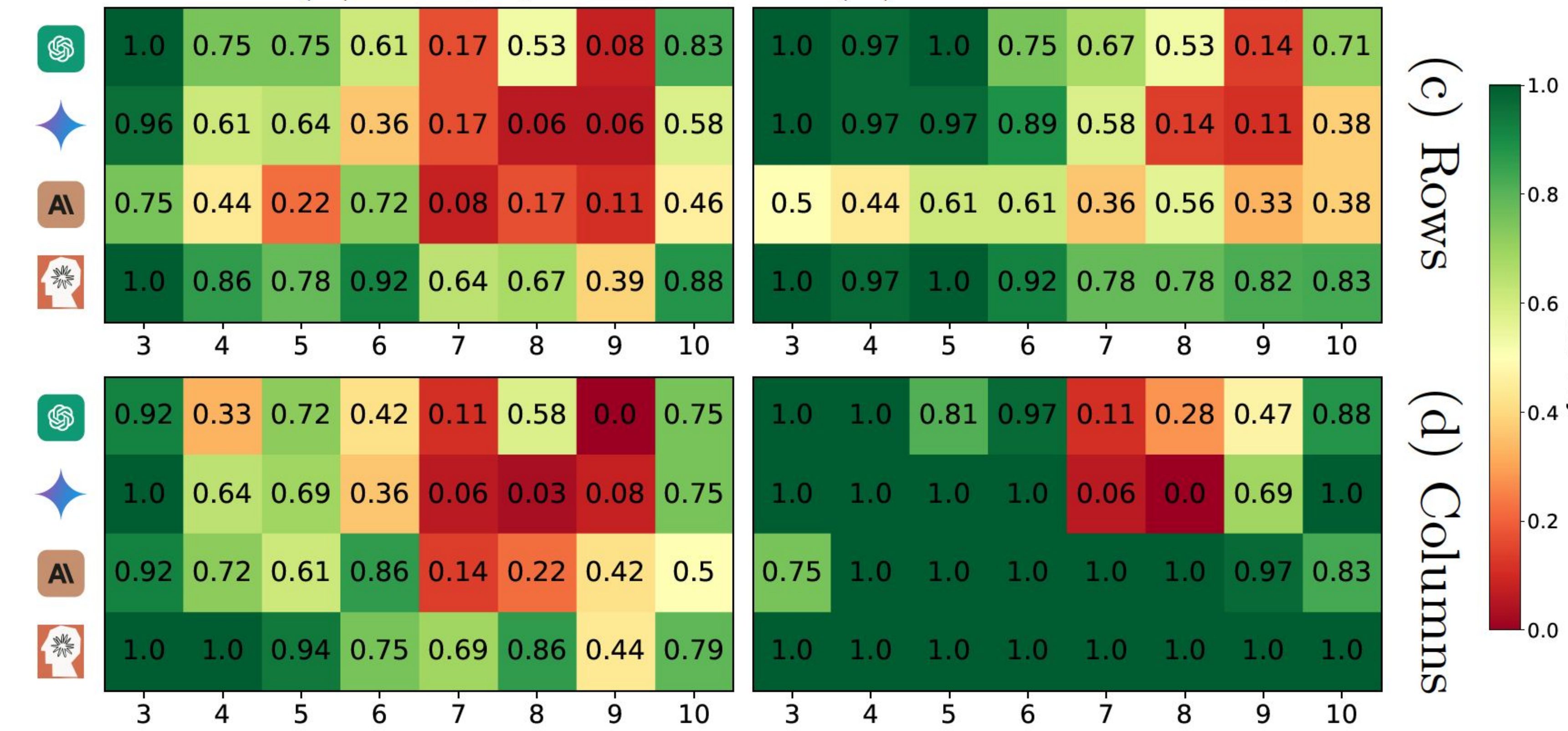
P1: How many times do the blue and red lines touch each other? Answer with a number in curly brackets, e.g., {5}.
P2: Are the two circles overlapping? Answer with Yes/No.
P3: Which character is being highlighted with a red oval? Please provide your answer in curly brackets, e.g. {a}
P4: How many circles are in the image? Answer with only the number in numerical format.
P5: How many squares are in the image? Please answer with a number in curly brackets e.g., {10}.
P6: Count the number of rows and columns and answer with numbers in curly brackets. For example, rows={5} columns={6}.
P7: How many single-color paths go from A to D? Answer with a number in curly brackets e.g. {3}.

VLMs **cannot** clearly see if two **circles overlap** or not



VLMs **cannot count** the **rows and columns** in a grid





When can VLMs understand low-level details?

VLMs can get near **100% accuracy** on the tasks when the shapes are more **separated.**