

A Appendix

A.1 K-shot evaluation

We conduct the K-shot evaluations with GPT-4V, where K ranges from 0 to 4. The demonstrations are annotated by humans. The testing data is a subset of the Raven dataset, from which we select examples of the simplest mode (20 examples). For each K, we run the evaluation 4 times. The results are shown in Figure 4.

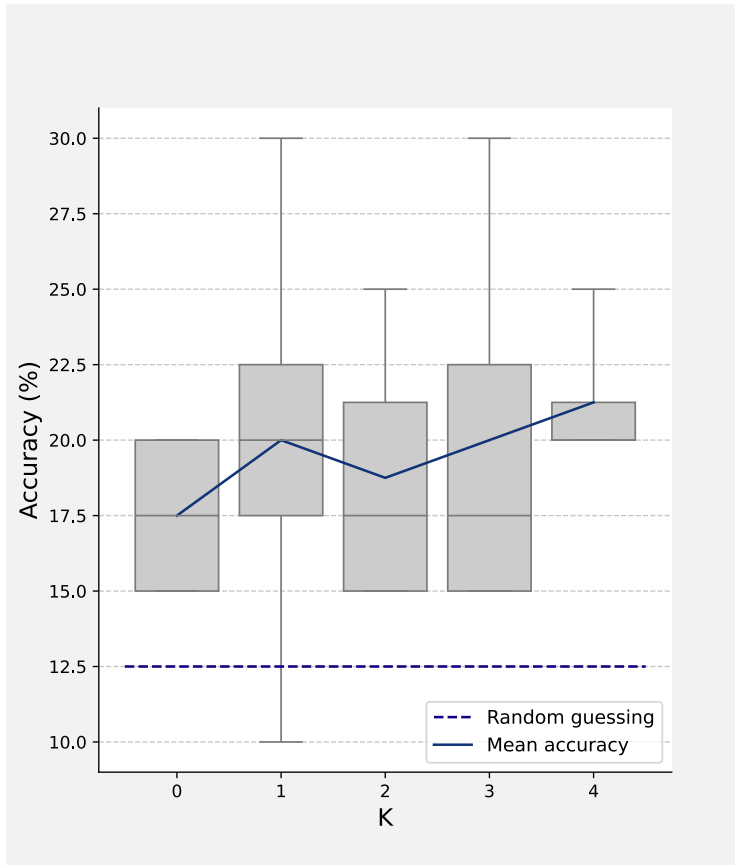


Figure 4: K-shot evaluation results with Raven subset.

From this figure, it is evident that few-shot evaluation generally outperforms zero-shot evaluation. However, when K is small, there is a large variance observed. As K increases, the mean accuracy slightly improves, and the variance significantly decreases. This trend is similar to observations in text-based large language models [Zhao et al. \(2021\)](#).

A.2 Prompts for different dataset and tasks

We listed all the prompts we used to generate the response from different models.

For the Mensa set we used the following prompt:

Mensa. (0-shot)

You can see a 3x3 grid of 9 boxes, one of which is empty (marked as ?). You have to choose which of the 6 alternative shapes (A-F) should be placed in the empty box in order to complete the pattern that connects the shapes. Finally, provide your prediction as Answer: "X"
{query image}

In Section 5.4, we explain why we chose the particular order of the prompt and query image order. We also provide the other prompts we used in below.

Mensa (1-shot)

You can see a 3x3 grid of 9 boxes, one of which is empty (marked as ?). You have to choose which of the 6 alternative shapes (A-F) to be placed in the empty box in order to complete the pattern that connects the shapes. Think step by step by first describe the each box in the 3x3 grid, and each of the alternative shapes as the candidate answers. Then identify the common pattern. Finally, provide your prediction as Answer: "X"

For example, for the following image:

{in-context image}

{in-context description}

{in-context rationale}

{in-context answer}

Now do the following one:

{query image}

Mensa (1-shot) + Oracle Desc.

You can see a 3x3 grid of 9 boxes, one of which is empty (marked as ?). You have to choose which of the 6 alternative shapes (A-F) to be placed in the empty box in order to complete the pattern that connects the shapes. Think step by step by first describe the each box in the 3x3 grid, and each of the alternative shapes as the candidate answers. Then identify the common pattern. Finally, provide your prediction as Answer: "X"

For example, for the following image:

{in-context image}

{in-context description}

{in-context rationale}

{in-context answer}

Now do the following one:

{query image}

{query oracle description}

Mensa (1-shot) + Oracle Desc. + Oracle Rationale

You can see a 3x3 grid of 9 boxes, one of which is empty (marked as ?). You have to choose which of the 6 alternative shapes (A-F) to be placed in the empty box in order to complete the pattern that connects the shapes. Think step by step by first describe the each box in the 3x3 grid, and each of the alternative shapes as the candidate answers. Then identify the common pattern. Finally, provide your prediction as Answer: "X"

For example, for the following image:

{in-context image}

{in-context description}

{in-context rationale}

{in-context answer}

Now do the following one:

{query image}

{query oracle description}

{query oracle rationale}

IntelligenceTest (0-shot)

This image is an Intelligence Test Question asking which figure (A-F) belongs in the bottom right box. Please select the correct answer. You must first give your explanation and then output the answer at the end of your response in the format: "The correct answer is: _".

{query image}

RAVEN (0-shot)

You can see a 3x3 grid of 9 boxes, one of which is empty (marked as ?). You have to choose which of the 8 alternative shapes (A-H) should be placed in the empty box in order to complete the pattern that connects the shapes. You must first give your explanation and then output the answer at the end of your response in the format: "The correct answer is: _".

{query image}

Segmented Prompt (0-shot)

In the first image, you will see a 3x3 grid of 9 boxes. Each row has three images and is to be read from left to right, with the last image in the third row is missing (marked as ?). Your task is to infer the correct pattern that should complete each row based on the sequence observed in the preceding patterns, and finally select the right option (A, B, C, D, E, F) that fits the 3rd row's last image.

{query image}

For your convenience, I provide 15 segmented figures: the question grid is segmented into 9 parts, and the answer options are segmented into 6 parts. q0, q1, and q2 are the first row, q3, q4, and q5 are the second row, and q6, q7, and q8 are the third row. A, B, C, D, E, and F are the answer options. Your task is to find which option should be placed in q8.

q0: {q0 image}

q1: {q1 image}

q2: {q2 image}

q3: {q3 image}

q4: {q4 image}

q5: {q5 image}

q6: {q6 image}

q7: {q7 image}

q8: {q8 image}

A: {A image}

B: {B image}

C: {C image}

D: {D image}

E: {E image}

F: {F image}

For each row, analyze the changes and relationships between the images. Consider the number of shapes, the types of shapes, their positions, the shading, and any other changes that occur from one pattern to the next. Once you have identified the rule or sequence that applies to the rows, select the option (A, B, C, D, E, F) that contains the pattern which correctly completes the third row sequence.

Please first give your explanation and then write the answer at the end of your response in the format: "The correct answer is: _".

A.3 Sampling methods for model evaluations

In the main text, we present evaluation results obtained by independently running each model 10 times and estimating their respective statistics. For GPT4-V, we found that using a zero temperature setting along with different random seeds effectively balances prediction accuracy with reasoning diversity. In the case of Gemini Pro Vision, the optimal performance is achieved at a temperature setting of 0.4. However, for the self-consistency evaluation, we increase the sampling temperature to a maximum of 0.7. This adjustment facilitates more varied predictions which are beneficial for majority-voting processes. For all other models, we observed that a lower temperature range, between 0 and 0.2, is necessary. This lower temperature helps the models better adhere to instructions and improves their task prediction accuracy, though it also results in somewhat less variability in the models' predictions.

M-Easy (Oracle)

Description

The grid:

1. Top left box: square with a cross sign.
2. Top center box: circle with a star.
3. Top right box: Empty triangle.
4. Middle left box: empty square.
5. Middle center box: circle with a cross sign.
6. Middle right box: triangle with a star.
7. Bottom left box: square with a star.
8. Bottom center box: empty circle.
9. Bottom right box: '?'.

The alternative shapes are:

- A. Triangle with a star.
- B. Triangle with a plus sign.
- C. Circle with a cross sign.
- D. Circle with a star.
- E. Empty triangle.
- F. Triangle with a cross sign.

Rationale

Analyzing the grid, there seems to be a pattern involving both the shapes and the symbols inside them:

- The symbols inside the shapes alternate between a cross, a star, and then a blank space as you move horizontally across each row.
- The shapes themselves also alternate within each row – this is seen with the square, circle, and triangle repeating in each row in that order.

A.4 Oracle description and rationale for each RPMs

M-Medium (Oracle)
<div data-bbox="329 569 1292 1352"> <p>Description</p> <p>The grid:</p> <ol style="list-style-type: none"> 1. Top left box: A downward-pointing triangle with three dots in it. 2. Top center box: A leftward-pointing triangle with two dots in it. 3. Top right box: An upward-pointing triangle with one dot in it. 4. Middle left box: A rightward-pointing triangle with two dots in it. 5. Middle center box: A downward-pointing triangle with one dot in it. 6. Middle right box: A leftward-pointing triangle with three dots in it. 7. Bottom left box: An upward-pointing triangle with one dot in it. 8. Bottom center box: A rightward-pointing triangle with three dots in it. 9. Bottom right box: '?'. <p>The alternative shapes are:</p> <ul style="list-style-type: none"> • A. An upward-pointing triangle with two dots in it. • B. A downward-pointing triangle with one dot in it. • C. A leftward-pointing triangle with one dot in it. • D. A rightward-pointing triangle with two dots in it. • E. A leftward-pointing triangle with two dots in it. • F. A downward-pointing triangle with two dots in it. </div> <div data-bbox="329 1373 1292 1902"> <p>Rationale</p> <p>Analyzing the grid, it appears that there's a pattern related to the direction the triangle is pointing and the number of dots within the triangles. First, let's establish the patterns of triangle directions and dots count:</p> <ul style="list-style-type: none"> • The first row has the triangles pointing downward, to the left, and then up. • The second row has the triangles pointing rightward, downward, and then to the left. • This implies that the direction that the triangle is pointing to is rotating clockwise in each row. <p>Now let's look at the pattern in the number of dots:</p> <ul style="list-style-type: none"> • The first row has 3, 2, 1 dots. • The second row has 2, 1, 3 dots. • This implies a pattern of a rotation of a decreasing sequence. </div>

Hard (Oracle)
<p>Description</p> <p>The grid:</p> <ol style="list-style-type: none"> 1. Top left box: white circle, white triangle, black square. 2. Top center box: white triangle, black circle, white square. 3. Top right box: black square, black square, white triangle. 4. Middle left box: black circle, white square, white triangle. 5. Middle center box: black square, white triangle, black square. 6. Middle right box: white triangle, black square, white circle. 7. Bottom left box: white triangle, black square, black square. 8. Bottom center box: black square, white circle, white triangle. 9. Bottom right box: '?'. <p>The alternative shapes are:</p> <ul style="list-style-type: none"> • A. white circle, white triangle, black square. • B. black circle, white square, white triangle. • C. white circle, white square, black triangle. • D. white circle, black square, white triangle. • E. black square, white triangle, white circle. • F. white square, white triangle, black circle.
<p>Rationale</p> <p>Analyzing the grid, it appears that there's a pattern along the diagonal lines from top right to bottom left direction.</p> <ul style="list-style-type: none"> • The three elements in the first diagonal line (top right, middle center, bottom left) consist of a black square, a black square, and a white triangle. As we follow the diagonal line, these elements rotate 120 degrees clockwise. • The three elements in the second diagonal line (top left, middle right, bottom center) consist of a white circle, a white triangle, and a black square. As we follow the diagonal line, these elements rotate 120 degrees clockwise. • This implies that the third diagonal line (top center, middle left, bottom right) should also follow this pattern.

A.5 Description generated by GPT-4V

We provide several additional sampled descriptions in the table below that illustrate the perceptual blind spot in describing RPM problems. GPT-4V encounters challenges in recognizing and differentiating intricate spatial relationships and patterns in the M-Medium and M-Hard problem sets.

M-Medium (description - additional samples #2)

The grid:

1. Top left box: Inverted triangle with three dots clustered towards the bottom.
2. Top center box: Upright triangle with three dots clustered towards the narrow point.
3. Top right box: Inverted triangle with one dot in the center.
4. Middle left box: Upright triangle with three dots forming a line along the base.
5. Middle center box: Inverted triangle with two dots, aligned vertically in the center.
6. Middle right box: Upright triangle with two dots, one at the tip and one at the center of the base.
7. Bottom left box: Inverted triangle with two dots forming a line along the base.
8. Bottom center box: "?".

The alternative shapes are:

- A. Upright triangle with two dots side by side at the base.
- B. Inverted triangle with one dot at the center of the base.
- C. Upright triangle with one dot in the center.
- D. Inverted triangle with two dots side by side near the tip.
- E. Upright triangle with two dots, one at the center of the base and one at the tip.
- F. Upright triangle with three dots, forming a line along the base.

M-Medium (description - additional samples #3)

The grid:

1. Top left box: An inverted triangle with three black dots forming a vertical line at the center.
2. Top center box: An upright triangle with three black dots forming a vertical line at the center.
3. Top right box: An inverted triangle with a single black dot at the center.
4. Middle left box: An upright triangle with two black dots forming a horizontal line at the center.
5. Middle center box: An inverted triangle with two black dots forming a horizontal line at the center.
6. Middle right box: An upright triangle with a single black dot at the center.
7. Bottom left box: An inverted triangle with one black dot at a bottom vertex.
8. Bottom center box: "?".

The alternative shapes are:

- A. An upright triangle with two black dots forming a vertical line at the center.
- B. An upright triangle with a single black dot at the center.
- C. An inverted triangle with a single black dot at a bottom vertex.
- D. An inverted triangle with two black dots forming a horizontal line at the center.
- E. An upright triangle with two black dots forming a horizontal line at the center.
- F. An inverted triangle with three black dots forming a horizontal line at the center.

M-Hard (description - additional samples #1)

The grid:

1. Top left box: hollow circle, hollow triangle, solid square.
2. Top center box: solid circle, hollow square, hollow triangle.
3. Top right box: solid square, solid triangle, hollow circle.
4. Middle left box: solid circle, hollow triangle, hollow square.
5. Middle center box: solid square, solid triangle, solid circle.
6. Middle right box: hollow square, hollow circle, solid triangle.
7. Bottom left box: hollow triangle, solid square, solid circle.
8. Bottom center box: solid circle, hollow triangle, hollow square.
9. Bottom right box: "?".

The alternative shapes are:

- A. Hollow circle, hollow triangle, solid square.
- B. Solid circle, hollow square, hollow triangle.
- C. Hollow circle, hollow square, solid triangle.
- D. Hollow circle, hollow triangle, solid square.
- E. Solid square, hollow triangle, hollow circle.
- F. Hollow square, hollow triangle, solid circle.

M-Hard (description - additional samples #2)

The grid:

1. Top left box: circle at the top, triangle in the middle, square at the bottom.
2. Top center box: triangle at the top, circle in the middle, small square at the bottom.
3. Top right box: square at the top, triangle in the middle, circle at the bottom.
4. Middle left box: circle at the top, square in the middle, triangle at the bottom.
5. Middle center box: square at the top, two triangles in the middle, circle at the bottom.
6. Middle right box: triangle at the top, circle in the middle, square at the bottom.
7. Bottom left box: triangle at the top, two squares in the middle, circle at the bottom.
8. Bottom center box: square at the top, circle in the middle, two triangles at the bottom.
9. Bottom right box: "?".

The alternative shapes are:

- A. Circle at the top, triangle in the middle, square at the bottom.
- B. Circle at the top, square in the middle, two triangles at the bottom.
- C. Circle at the top, small square in the middle, triangle at the bottom.
- D. Two circles at the top, square in the middle, triangle at the bottom.
- E. Square at the top, circle in the middle, triangle at the bottom.
- F. Triangle at the top, square in the middle, circle at the bottom.

A.6 Rationale generated by GPT-4V

We present several examples that demonstrate the reasoning capabilities of GPT-4V for solving certain RPM tasks. From the examples, it becomes evident that as the difficulty level of the problem increases, the GPT-4V tends to struggle with its reasoning abilities and provides an explanation that is nonsensical.

Generated rationale for *M-Medium* RPM

Analyzing the grid, it appears there's a pattern related to the direction the triangle is pointing and the number of dots within the triangles.

First, let's establish the patterns of triangle directions and dots count:

- The first row has the triangles pointing downward, to the left, and then up.
- The second row has the triangles pointing rightward, downward, and then to the left.
- It implies that for the third row, following the same pattern, the triangles should point upward, rightward, and then downward.

Now let's look at the pattern in the number of dots:

- Along the vertical columns, we can observe a pattern in dot counts:
 - The first column has 3, 2, 1 dots.
 - The second column has 2, 1, 3 dots.
 - If we follow this pattern, the third column should have 1, 3, 2 dots from top to bottom.

Given these observed sequences, the third row should have a triangle pointing downward with two dots in it, fitting the patterns of both direction and dot count. Looking at the alternative shapes, option F is the downward-pointing triangle with two dots within it.

Answer: "F"

Generated rationale for *M-Hard* RPM

Analyzing the grid, we can deduce the following pattern:

1. The shapes in each box follow a sequence that changes from one cell to the next horizontally.
2. There are three types of shapes: circle, square, and triangle, and each type also has two variants of color: black or white.

To understand how the sequence works, we can lay out the patterns we see:

- In the top row, we start with a white circle, white triangle, and black square. Moving right, the shapes rotate positions clockwise, and their colors invert (the white circle becomes a black circle and so on).
- The same pattern follows in the middle row: black circle, white square, white triangle becomes black square, white triangle, black square, and again inverting colors and rotating the shapes clockwise.
- Following this pattern in the bottom row, white triangle, black square, black square should change by rotating the shapes clockwise and inverting the colors. The black square would become a white square, the black square would become a white circle, and the white triangle would stay the same, remaining in the bottom position.

Matching this description with the alternatives given, option "F" (white square, white triangle, black circle) fits the observed sequence.

Answer: "F"