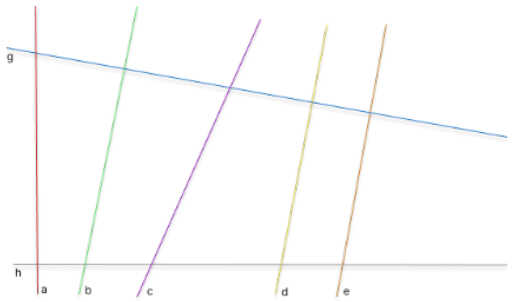
In this document, the tasks were graded. After grading every task in every mechanism, the analysis of finding an average grade for all tasks was done, as well as an average grade of mechanisms and modes. There are some comments written, that were used for writing later the Master thesis and qualitative answers.

1. **Which of the lines are perpendicular to each other? Check with the triangle ruler. Example: a ⊥ g. Please solve this step by step.**



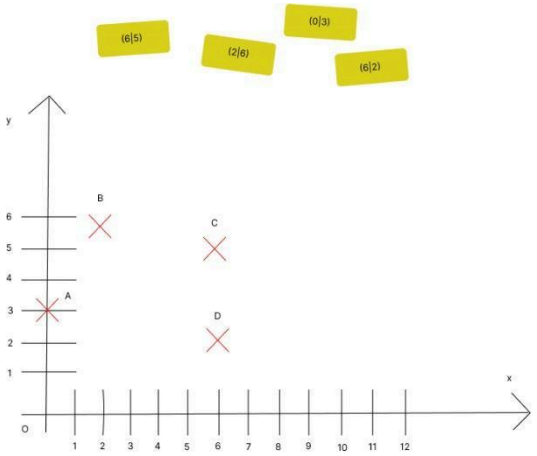
First prompting:
A lot of explanations, but not correct – zero shot -
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
A lot of explanations, but not correct – few shot
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
Chain-of-thought – symbolic - not correct
Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 1.66**
Chain-of-thought – commonsense - not correct
Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 1.66**
Chain-of-thought – arithmetic - not correct, not even diff. a grey line
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Every mechanism in one window prompted for one task each
Zero-shot
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 2**

Few-shot
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 2**
Commonsense
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 2**
Arithmetic
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 2**
Symbolic
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 2**
11 years old pupil prompt Commonsense
not correct, short explanation

Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 1.66**
11 years old pupil prompt - Arithmetic
Good general explanation, without a final answer, not specific, not correct answers at the end
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 2.66**
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 2;
11 years old pupil prompt - Symbolic
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**

11 years old pupil prompt zero shot- a lot of explanations, but not correct
Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 1.66**
11 years old pupil prompt Few-shot
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**

2. Arrange the cards to the shown dots.



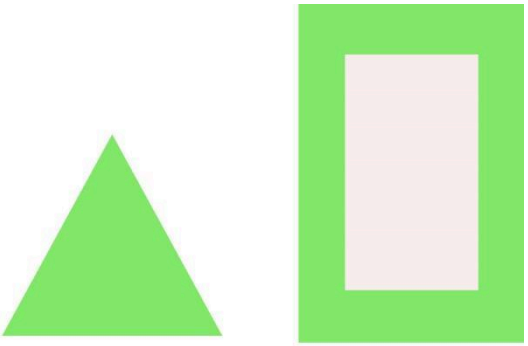
11 years old pupil prompt zero shot
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
First prompting – few shot

Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**

Chain-of-thought – symbolic - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Chain-of-thought – commonsense - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Chain-of-thought – arithmetic - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Every mechanism in one window prompted for one task each – correct in every mechanism

Commonsense
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Zero-shot
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Few-shot
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Symbolic
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Arithmetic
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
11 years old pupil prompt
Commonsense
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
11 years old pupil prompt Arithmetic – 1 not correct, 3 correct yellow cards
Accuracy grade: 2; Alignment with a task: 3; Age-Appropriate - 2; **Average grade: 2.33**
11 years old pupil prompt Symbolic – 1 not correct, 3 correct yellow cards
Accuracy grade: 2; Alignment with a task: 3; Age-Appropriate - 2; **Average grade: 2.33**
11 years old pupil prompt zero shot
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
11 years old pupil prompt Few-shot
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**

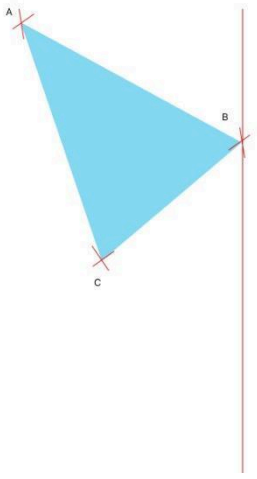
3. Transfer the figures into your notebook and mark the axes of symmetry.



Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
First prompting – zero shot - half correct explanation, half not, NO PICTURE
Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**
First prompting – few shot - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Chain-of-thought – symbolic - Rectangle RIGHT, Triangle not
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**

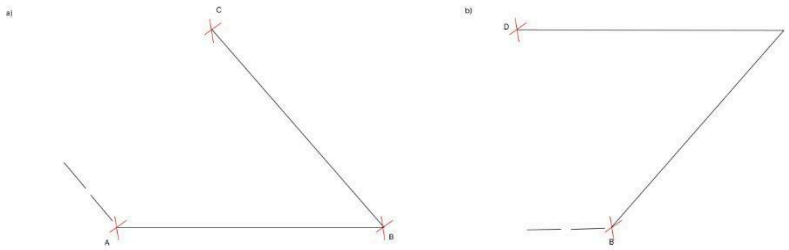
Chain-of-thought – commonsense – correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Chain-of-thought – arithmetic - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Every task and every mechanism in one chatGPT4 chat
Zero-shot
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Few-shot
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Arithmetic
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Symbolic
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Commonsense
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
11 years old pupil prompt Commonsense - not correct explanation for a triangle, correct explanation for rectangle, but no picture
Accuracy grade: 2; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 2.66**
11 years old pupil prompt Arithmetic - not correct explanation for a triangle, correct explanation for rectangle, but no picture
Accuracy grade: 2; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 2.66**
11 years old pupil prompt Symbolic – correct explanation, no picture
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old pupil prompt zero shot - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
11 years old pupil prompt Few-shot - not correct explanation for a triangle, correct explanation for rectangle, but no picture

4. Complete the figure to make it axially symmetrical. The red line is the axis of symmetry.



First prompting – zero shot - not correct
Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 1.66**
First prompting – Few shot-
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
Chain-of-thought – symbolic
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Chain-of-thought – commonsense - not correct
Chain-of-thought – arithmetic - Correct explanation, NO PICTURE
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**
Every task and every mechanism in one chatGPT4 chat – arithmetic - Correct explanation, NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
Zero shot
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Few shot
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Symbolic
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Commonsense
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
artihemthic
11 years old pupil prompt Commonsense - Correct explanation, NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old pupil prompt Arithmetic - Correct explanation, NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old pupil prompt Symbolic – Correct explanation, NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old prompt zero shot - Correct explanation, NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old prompt Few-shot - Good explanation, NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**

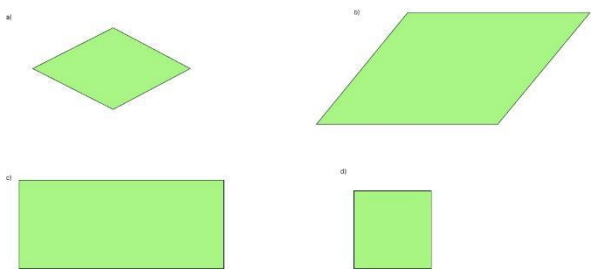
5. Transfer the figure into your notebook and complete it to form a parallelogram.



First prompting – zero shot - not correct
Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 1.33**
First prompting - Few-shot

The instruction is ok, BUT NO PICTURE AT ALL
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 2**
Chain-of-thought – symbolic- not correct
Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 1.33**
Chain-of-thought – commonsense - One part of the response is correct, it made a picture for one parallelogram but not correct picture
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2**
Chain-of-thought – arithmetic - One part of the response is correct, another not, NO PICTURE
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2**
Every task and every mechanism in one chatGPT4 chat – zero shot - correct explanation, no picture
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
Few shot not correct
Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 1.33**
Commonsense,
Accuracy grade: 2; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 2.66**
symbolic
Accuracy grade: 2; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 2.66**
Arithmetic
Accuracy grade: 2; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 2.66**
11 years old pupil prompt Commonsense - Correct explanation, NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old pupil prompt Arithmetic - Correct explanation for a) left parallelogram, not for b) second one NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old pupil prompt Symbolic - Correct explanation for a) left parallelogram, not for b) second one NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old pupil prompt zero shot - Correct explanation for a) left parallelogram, not for b) second one NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old pupil prompt Few-shot - Correct explanation, NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**

6. Which quadrilaterals are rhombuses?

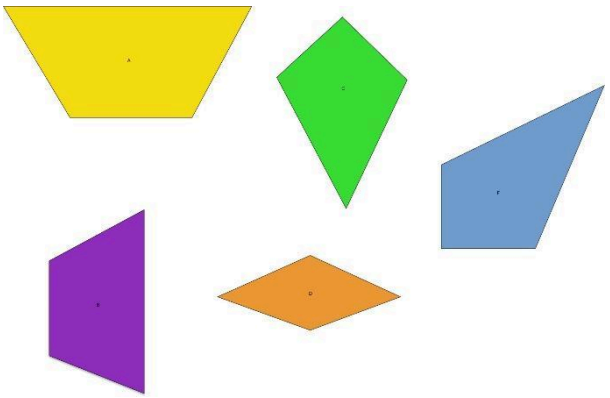


First prompting – zero shot

Correct answer. But did not have further explanation for the other ones
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
First prompting - Few-shot
Correct answer. But did not have further explanation for the other ones
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
Chain-of-thought – symbolic - not correct
Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 1.33**
Chain-of-thought – commonsense - Correct answer
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Chain-of-thought – arithmetic - Correct answer
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Every mechanism in one window prompted for one task each – zero shot
- Correct answer in every mechanism
few shot; commonsense, symbolic, arithmetic

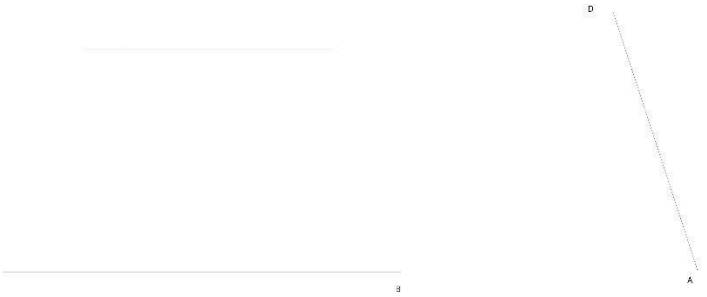
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
11 years old pupil prompt Commonsense - Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
11 years old pupil prompt Arithmetic - Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
11 years old pupil prompt Symbolic – Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
11 years old pupil prompt zero shot - Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
11 years old pupil prompt Few-shot - Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**

7. Which of these quadrilaterals are:
   a) Kites b) Trapezoids



First prompting – zero shot - one part correct
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
First prompting - Few-shot - not correct
Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate - 2; **Average grade: 1.66**
Chain-of-thought – symbolic - not correct fully, BLUE kite it's seen as a trapezoid
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
Chain-of-thought – commonsense - not correct fully, BLUE kite it's seen as a trapezoid
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
Chain-of-thought – arithmetic - not correct fully, BLUE kite it's seen as a trapezoid
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
Every mechanism in one chatGPT4 chat prompted for one task each – zero shot - not correct fully
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
Few-shot
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
BLUE kite it's seen as a trapezoid and purple is seen as not belonging anywhere in commonsense
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
symbolic
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
arithmetic
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
commonsense
11 years old pupil prompt Commonsense - not correct fully, BLUE kite it's seen as a trapezoid, purple has seen neither as a trapezoid neither as a kite
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
11 years old pupil prompt Arithmetic - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
11 years old pupil prompt Symbolic – 4 correct, one not
Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**
11 years old pupil prompt zero shot - not correct fully, BLUE kite it's seen as a trapezoid, purple has seen neither as a trapezoid neither as a kite
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
11 years old pupil prompt Few-shot - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**

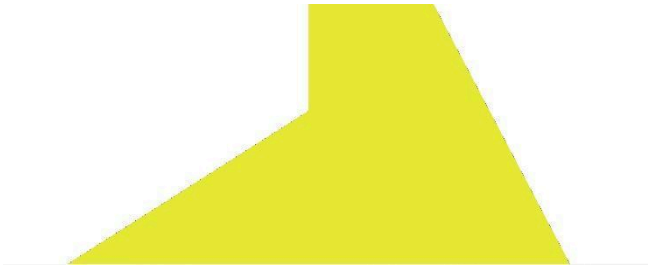
8. Transfer the figure to your notebook and complete it to form a symmetrical trapezoid. Mark all the corner points and measure the side lengths.



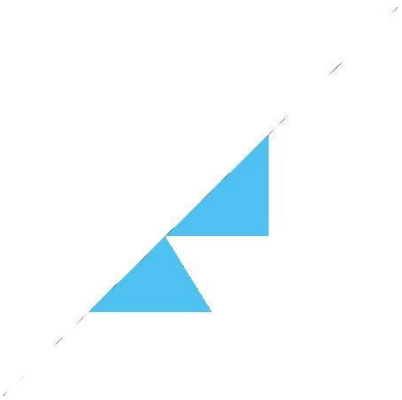
First prompting – zero shot - Good instructions, NO PICTURE
Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**
First prompting - Few-shot
Not correct, no picture, it said it drew the picture, but it did not explain
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Chain-of-thought – symbolic- not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
Chain-of-thought – commonsense - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
Chain-of-thought – arithmetic - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
Every mechanism in one window prompted for one task each –
one shot
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
few shot
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
symbolic
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
commonsense
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
arithemtic Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
11 years old pupil prompt Commonsense – it is kinda correct, but it's not understandable which trapezoid chatGPT4 is pointing to, probably for the one on the right side
Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**
11 years old pupil prompt Arithmetic - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
11 years old pupil prompt Symbolic – not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
11 years old pupil prompt zero shot - Good instructions for one parallelogram – the right one, NO PICTURE, no measures

Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
11 years old pupil prompt Few-shot - Good instructions for one parallelogram – the right one, NO PICTURE
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**

9. Complete an axially symmetrical figure.

a) 

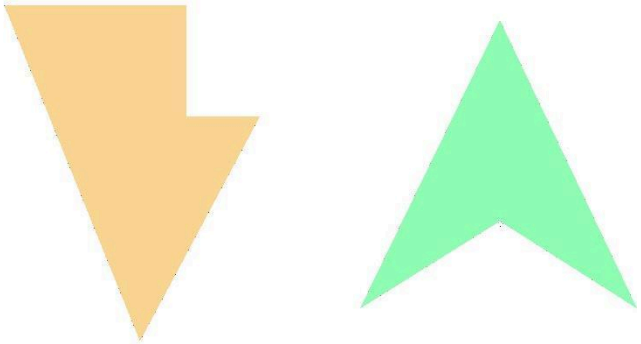
b)



First prompting – zero shot - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
First prompting - Few-shot
Not being able to access the files
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
Chain-of-thought – symbolic - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
Chain-of-thought – commonsense - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
Chain-of-thought – arithmetic - Correct INSTRUCTION, NO PICTURE
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
Every mechanism in one window prompted for one task each –
few shot
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**

commonsense
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
symbolic
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
Arithmetic
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
Zero shot
Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.33**
11 years old pupil prompt Commonsense – Correct INSTRUCTION, NO
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
PICTURE
11 years old pupil prompt Arithmetic - Correct INSTRUCTION, NO
PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old pupil prompt Symbolic – Correct INSTRUCTION, NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old pupil prompt zero shot - Correct INSTRUCTION, NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**
11 years old pupil prompt Few-shot - Correct INSTRUCTION, NO PICTURE
Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate - 3; **Average grade: 3**

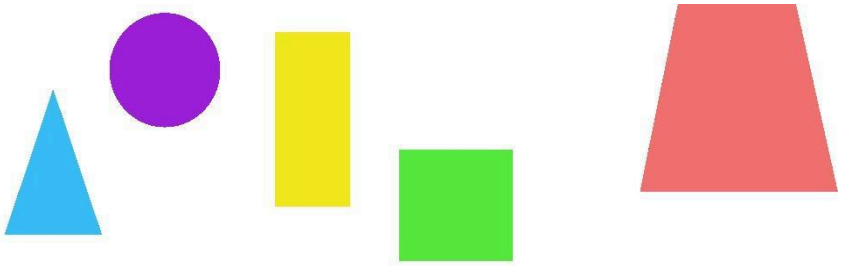
A) B) Complete a) to an axially symmetric figure and b) to a point-symmetric figure.



First prompting – zero shot - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
First prompting - Few-shot
Not correct, not being able to design the picture
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
Chain-of-thought – symbolic - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**

Chain-of-thought – commonsense - not correct
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
Chain-of-thought – arithmetic - Correct INSTRUCTION for green figure, not for orange
Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**
Every mechanism in one window prompted for one task each – few shot
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
commonsense
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
Symbolic
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 1; **Average grade: 1**
zero shot
Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**
Arithmetic – long, complex, non-understandable explanation
Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate - 2; **Average grade: 1.33**
11 years old pupil prompt Commonsense – Correct INSTRUCTION for green figure not for orange, NO PICTURE
Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**
11 years old pupil prompt Arithmetic - Correct INSTRUCTION for green figure not for orange, NO PICTURE
Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**
11 years old pupil prompt Symbolic - Correct INSTRUCTION for green figure not for orange, NO PICTURE
Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**
11 years old pupil prompt zero shot - Correct INSTRUCTION for green figure not for orange, NO PICTURE
Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**
11 years old pupil prompt Few-shot - Correct INSTRUCTION for green figure not for orange, NO PICTURE
Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate - 3; **Average grade: 2.66**

10. Give the names of the characters.



First prompting – zero shot - Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
First prompting - Few-shot- Correct

Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Chain-of-thought – symbolic- Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Chain-of-thought – commonsense – Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Chain-of-thought – arithmetic - Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
Every mechanism in one window prompted for one task each – Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
11 years old pupil prompt Commonsense – Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
11 years old pupil prompt Arithmetic – Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
11 years old pupil prompt Symbolic – Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
11 years old pupil prompt zero shot - Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**
11 years old pupil prompt Few-shot - Correct
Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate - 4; **Average grade: 4**

1 task: 1.86 average score grade

2. task: 2.35 average score grade

3 task: 2.50 average score grade

4 task: 2.21 average score grade

5 task: 2.39 average score grade

6 task: 3.68 average score grade

7 task: 1.97 average score grade

8 task: 1.35 average score grade

9 task A): 1.93 average score grade

9 task B): 1.84 average score grade

10 task: 4 average score grade

Total average score grade: 2.6

1. task

First prompting – zero shot

Zero Shot:

- Accuracy grade: 2

- Alignment with task: 2

- Age-Appropriate: 3

- Average grade: 2.33

First prompting – few shot

Few-shot:

- Accuracy grade: 2

- Alignment with task: 2

- Age-Appropriate: 3

- Average grade: 2.33

Chain-of-thought – symbolic

Symbolic:

- Accuracy grade: 1

- Alignment with task: 2

- Age-Appropriate: 2

- Average grade: 1.66

Chain-of-thought – commonsense

Commonsense:

- Accuracy grade: 1

- Alignment with task: 2

- Age-Appropriate: 2

- Average grade: 1.66

Chain-of-thought – arithmetic

Arithmetic:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

Every mechanism in one window prompted for one task each

Zero-shot:

- Accuracy grade: 2

- Alignment with task: 2

- Age-Appropriate: 2

- Average grade: 2

Few-shot:

- Accuracy grade: 2

- Alignment with task: 2

- Age-Appropriate: 2

- Average grade: 2

Commonsense:

- Accuracy grade: 2

- Alignment with task: 2

- Age-Appropriate: 2

- Average grade: 2

Arithmetic:

- Accuracy grade: 2

- Alignment with task: 2

- Age-Appropriate: 2

- Average grade: 2

Symbolic:

- Accuracy grade: 2
- Alignment with task: 2
- Age-Appropriate: 2
- Average grade: 2

11 years old pupil prompt Commonsense

Commonsense:

- Accuracy grade: 1
- Alignment with task: 2
- Age-Appropriate: 2
- Average grade: 1.66

11 years old pupil prompt - Arithmetic

Arithmetic:

- Accuracy grade: 1
- Alignment with task: 1
- Age-Appropriate: 2
- Average grade: 1.33

11 years old pupil prompt - Symbolic

Symbolic:

- Accuracy grade: 1
- Alignment with task: 1
- Age-Appropriate: 2
- Average grade: 1.33

11 years old pupil prompt zero shot

Zero Shot:

- Accuracy grade: 1
- Alignment with task: 2
- Age-Appropriate: 2
- Average grade: 1.66

11 years old pupil prompt Few-shot

Few-shot:

- Accuracy grade: 1
- Alignment with task: 1
- Age-Appropriate: 2
- Average grade: 1.33

Summary

Zero Shot:

- Average grade: (2.33 + 2 + 1.66) / 3 = 1.99

Few-shot:

- Average grade: (2.33 + 2 + 1.33) / 3 = 1.75

Symbolic:

- Average grade: (1.66 + 2 + 1.33 + 1.33) / 4 = 1.66

Commonsense:

- Average grade: (1.66 + 2 + 1.66) / 3 = 1.77

Arithmetic:

- Average grade: (1.33 + 2 + 1.33 + 1.33) / 4 = 1.50

2. task

11 years old pupil prompt zero shot

Zero Shot:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

First prompting – few shot

Few-shot:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

Chain-of-thought – symbolic

Symbolic:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

Chain-of-thought – commonsense

Commonsense:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

Chain-of-thought – arithmetic

Arithmetic:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

Every mechanism in one window prompted for one task each

Commonsense:

- Accuracy grade: 4

- Alignment with task: 4

- Age-Appropriate: 4

- Average grade: 4

Zero Shot:

- Accuracy grade: 4

- Alignment with task: 4

- Age-Appropriate: 4

- Average grade: 4

Few-shot:

- Accuracy grade: 4

- Alignment with task: 4

- Age-Appropriate: 4

- Average grade: 4

Symbolic:

- Accuracy grade: 4

- Alignment with task: 4

- Age-Appropriate: 4

- Average grade: 4

Arithmetic:

- Accuracy grade: 4

- Alignment with task: 4

- Age-Appropriate: 4

- Average grade: 4

11 years old pupil prompt Commonsense

Commonsense:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

11 years old pupil prompt Arithmetic

Arithmetic:

- Accuracy grade: 2

- Alignment with task: 3

- Age-Appropriate: 2

- Average grade: 2.33

11 years old pupil prompt Symbolic

Symbolic:

- Accuracy grade: 2

- Alignment with task: 3

- Age-Appropriate: 2

- Average grade: 2.33

11 years old pupil prompt zero shot

Zero Shot:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

11 years old pupil prompt Few-shot

Few-shot:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

Summary

Zero Shot:

- Average grade: (1.33 + 4 + 1.33) / 3 = 2.22

Few-shot:

- Average grade: (1.33 + 4 + 1.33) / 3 = 2.22

Symbolic:

- Average grade: (1.33 + 4 + 2.33) / 3 = 2.55

Commonsense:

- Average grade: (1.33 + 4 + 1.33) / 3 = 2.22

Arithmetic:

- Average grade: (1.33 + 4 + 2.33) / 3 = 2.55

3. Task

11 years old pupil prompt zero shot

Zero Shot:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

First prompting – zero shot

Symbolic:

- Accuracy grade: 3

- Alignment with task: 2

- Age-Appropriate: 3

- Average grade: 2.66

First prompting – few shot

Symbolic:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

Chain-of-thought – symbolic

Symbolic:

- Accuracy grade: 2

- Alignment with task: 2

- Age-Appropriate: 3

- Average grade: 2.66

Chain-of-thought – commonsense

Commonsense:

- Accuracy grade: 4
- Alignment with task: 4
- Age-Appropriate: 4
- Average grade: 4

Chain-of-thought – arithmetic

Arithmetic:

- Accuracy grade: 1
- Alignment with task: 1
- Age-Appropriate: 2
- Average grade: 1.33

Every task and every mechanism in one chatGPT4 chat

Zero-shot:

- Accuracy grade: 4
- Alignment with task: 4
- Age-Appropriate: 4
- Average grade: 4

Few-shot:

- Accuracy grade: 4
- Alignment with task: 4
- Age-Appropriate: 4
- Average grade: 4

Arithmetic:

- Accuracy grade: 1
- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

Symbolic:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

Commonsense:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

11 years old pupil prompt Commonsense

Commonsense:

- Accuracy grade: 2

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 2.66

11 years old pupil prompt Arithmetic

Arithmetic:

- Accuracy grade: 2

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 2.66

11 years old pupil prompt Symbolic

Symbolic:

- Accuracy grade: 3

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 3

11 years old pupil prompt Few-shot

Few-shot:

- Accuracy grade: 2

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 2.66

Summary

Zero Shot:

- Average grade: (1.33 + 1.33 +3) / 3 = 1.88

Few-shot:

- Average grade: (1.33 + 4 + 2.66) / 3 = 2.66

Symbolic:

- Average grade: (1.33 + 2.66 + 1.33) / 3 = 1.77

Commonsense:

- Average grade: (2.66 + 1.33 + 2.66) / 3 = 1.77

Arithmetic:

- Average grade: (2.66 + 1.33 + 1.33) / 3 = 1.77

4. task

First prompting – zero shot

Zero Shot:

- Accuracy grade: 1

- Alignment with task: 2

- Age-Appropriate: 2

- Average grade: 1.66

First prompting – few shot

Few-shot:

- Accuracy grade: 3

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 3

Chain-of-thought – symbolic

Symbolic:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

Chain-of-thought – commonsense

Commonsense:

- Accuracy grade: 1

- Alignment with task: 1

- Age-Appropriate: 2

- Average grade: 1.33

Chain-of-thought – arithmetic

Arithmetic:

- Accuracy grade: 2
- Alignment with task: 2
- Age-Appropriate: 3
- Average grade: 2.66

Every task and every mechanism in one chatGPT4 chat – arithmetic

Arithmetic:

- Accuracy grade: 3
- Alignment with task: 3
- Age-Appropriate: 3
- Average grade: 3

Every task and every mechanism in one chatGPT4 chat

Zero-shot:

- Accuracy grade: 3
- Alignment with task: 3
- Age-Appropriate: 3
- Average grade: 3

Few-shot:

- Accuracy grade: 1
- Alignment with task: 1
- Age-Appropriate: 2
- Average grade: 1.33

Arithmetic:

- Accuracy grade: 3
- Alignment with task: 3
- Age-Appropriate: 3

- Average grade: 3

Symbolic:

- Accuracy grade: 1
- Alignment with task: 1
- Age-Appropriate: 2
- Average grade: 1.33

11 years old pupil prompt Commonsense

Commonsense:

- Accuracy grade: 3
- Alignment with task: 3
- Age-Appropriate: 3
- Average grade: 3

11 years old pupil prompt Arithmetic

Arithmetic:

- Accuracy grade: 3
- Alignment with task: 3
- Age-Appropriate: 3
- Average grade: 3

11 years old pupil prompt Symbolic

Symbolic:

- Accuracy grade: 3
- Alignment with task: 3
- Age-Appropriate: 3
- Average grade: 3

11 years old pupil prompt zero shot

Zero Shot:

- Accuracy grade: 3

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 3

11 years old pupil prompt Few-shot

Few-shot:

- Accuracy grade: 3

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 3

Summary

Zero Shot:

- Average grade: (1.66 + 3 + 3) / 3 = 2.55

Few-shot:

- Average grade: (3 + 1.33 + 3) / 3 = 2.44

Symbolic:

- Average grade: (3 + 1.33 + 3) / 3 = 2.44

Commonsense:

- Average grade: (1.33 + 3 + 3) / 3 = 1.88

Arithmetic:

- Average grade: (2.66 + 3 + 1.33) / 3 = 2.88

5. task

First prompting – zero shot

Zero Shot:

- Accuracy grade: 1

- Alignment with task: 2

- Age-Appropriate: 2

- Average grade: 1.66

First prompting - Few-shot

Few-shot:

- Accuracy grade: 2

- Alignment with task: 2

- Age-Appropriate: 2

- Average grade: 2

Chain-of-thought – symbolic

Symbolic:

- Accuracy grade: 1

- Alignment with task: 2

- Age-Appropriate: 2

- Average grade: 1.66

Chain-of-thought – commonsense

Commonsense:

- Accuracy grade: 2

- Alignment with task: 2

- Age-Appropriate: 3

- Average grade: 2.33

Chain-of-thought – arithmetic

Arithmetic:

- Accuracy grade: 2
- Alignment with task: 2
- Age-Appropriate: 3
- Average grade: 2.33

Every task and every mechanism in one chatGPT4 chat – zero shot

Zero Shot:

- Accuracy grade: 3
- Alignment with task: 3
- Age-Appropriate: 3
- Average grade: 3

Every task and every mechanism in one chatGPT4 chat – few shot

Few-shot:

- Accuracy grade: 1
- Alignment with task: 2
- Age-Appropriate: 2
- Average grade: 1.66

Every task and every mechanism in one chatGPT4 chat – Commonsense

Commonsense:

- Accuracy grade: 2
- Alignment with task: 3
- Age-Appropriate: 3
- Average grade: 2.66

Every task and every mechanism in one chatGPT4 chat – Symbolic

Symbolic:

- Accuracy grade: 2

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 2.66

Every task and every mechanism in one chatGPT4 chat – Arithmetic

Arithmetic:

- Accuracy grade: 2

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 2.66

11 years old pupil prompt Commonsense

Commonsense:

- Accuracy grade: 3

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 3

11 years old pupil prompt Arithmetic

Arithmetic:

- Accuracy grade: 3

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 3

11 years old pupil prompt Symbolic

Symbolic:

- Accuracy grade: 3

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 3

11 years old pupil prompt Zero Shot

Zero Shot:

- Accuracy grade: 3

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 3

11 years old pupil prompt Few-shot

Few-shot:

- Accuracy grade: 3

- Alignment with task: 3

- Age-Appropriate: 3

- Average grade: 3

Summary

Zero Shot:

- Average grade: (1.66 + 3 + 3) / 3 = 2.56

Few-shot:

- Average grade: (2 + 1.66 + 3) / 3 = 2.22

Symbolic:

- Average grade: (1.66 + 2.66 + 3) / 3 = 2.44

Commonsense:

- Average grade: (2.33 + 2.66 + 3) / 3 = 2.66

Arithmetic:

- Average grade: (2.33 + 2.66 + 3) / 3 = 2.66

6. task

Zero Shot

1. **First prompting – zero shot**

   - Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate: 3; Average grade: 3

2. **Every mechanism in one window prompted for one task each – zero shot, few shot, commonsense symbolic, arithmetic**

   - (5 instances with Average grade: 4)

   - Total grade: \(4+4+4+4+4 = 20\)

   - Average grade: \(20 / 5 = 4\)

3. **11 years old pupil prompt – zero shot**

   - Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Few Shot

1. **First prompting – Few-shot**

   - Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate: 3; Average grade: 3

2. **11 years old pupil prompt – Few-shot**

   - Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Arithmetic

1. **Chain-of-thought – arithmetic**

   - Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

2. **11 years old pupil prompt – Arithmetic**

   - Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Commonsense

1. **Chain-of-thought – commonsense**

   - Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

2. **11 years old pupil prompt – Commonsense**

   - Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Symbolic

1. **Chain-of-thought – symbolic**

   - Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate: 2; Average grade: 1.33

2. **11 years old pupil prompt – Symbolic**

   - Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

3. Every mechanism and every task in oneChatGPT4 chat
   - Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4 every mechanism

Summary

- **Zero Shot:**

  - First prompting: Average grade: 3

  - Every mechanism in one window: Average grade: 4

  - 11 years old pupil: Average grade: 4

- **Few Shot:**

  - First prompting: Average grade: 3

  - 11 years old pupil: Average grade: 4

- **Arithmetic:**

  - Chain-of-thought: Average grade: 4

- 11 years old pupil: Average grade: 4

- **Commonsense:**
  - Chain-of-thought: Average grade: 4
  - 11 years old pupil: Average grade: 4

- **Symbolic:**
  - Chain-of-thought: Average grade: 1.33
  - 11 years old pupil: Average grade: 4

7.task

Zero Shot

First prompting mode:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

Every task and every mechanism in one chatGPT4 chat:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

11 years old pupil mode:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

Few Shot

First prompting mode:

- Accuracy grade: 1; Alignment with a task: 2; Age-Appropriate: 2; Average grade: 1.66

Few-shot (general):

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

11 years old pupil mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 2; Average grade: 1.33

Symbolic

First prompting mode:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

Every task and every mechanism in one chatGPT4 chat:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 2; Average grade: 1.33

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.66

Arithmetic

First prompting mode:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

Every task and every mechanism in one chatGPT4 chat:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 2; Average grade: 1.33

11 years old pupil mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 2; Average grade: 1.33

Commonsense

First prompting mode:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

Every task and every mechanism in one chatGPT4 chat:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 2; Average grade: 1.33

11 years old pupil mode:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

Summary

- **Zero Shot:**
  - First prompting mode: 2.33
  - Every task and every mechanism in one chatGPT4 chat: 2.33
  - 11 years old pupil mode: 2.33
  - **Overall Zero Shot Average:** \((2.33 + 2.33 + 2.33) / 3 = 2.33\)

- **Few Shot:**
  - First prompting mode: 1.66
  - General few-shot: 2.33
  - 11 years old pupil mode: 1.33
  - **Overall Few Shot Average:** \((1.66 + 2.33 + 1.33) / 3 = 1.77\)

- **Symbolic:**
  - First prompting mode: 2.33
  - Every task and every mechanism in one chatGPT4 chat: 1.33
  - 11 years old pupil mode: 2.66
  - **Overall Symbolic Average:** \((2.33 + 1.33 + 2.66) / 3 = 2.11\)

- **Arithmetic:**
  - First prompting mode: 2.33
  - Every task and every mechanism in one chatGPT4 chat: 1.33
  - 11 years old pupil mode: 1.33
  - **Overall Arithmetic Average:** \((2.33 + 1.33 + 1.33) / 3 = 1.66\)

- **Commonsense:**
  - First prompting mode: 2.33
  - Every task and every mechanism in one chatGPT4 chat: 1.33

- 11 years old pupil mode: 2.33

- **Overall Commonsense Average:** \((2.33 + 1.33 + 2.33) / 3 = 1.99\)

8. task

Zero Shot

First prompting mode:

- Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.66

Every task and every mechanism in one window:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

11 years old pupil mode:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

Few Shot

First prompting mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 2; Average grade: 1.33

Every task and every mechanism in one window:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

11 years old pupil mode:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

Symbolic

First prompting mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

Every task and every mechanism in one window:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

11 years old pupil mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

Arithmetic

First prompting mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

Every task and every mechanism in one window:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

11 years old pupil mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

Commonsense

First prompting mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

Every task and every mechanism in one window:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.66

Summary

- **Zero Shot:**
  - First prompting mode: 2.66
  - Every task and every mechanism in one window: 1
  - 11 years old pupil mode: 2.33
  - **Overall Zero Shot Average:** \((2.66 + 1 + 2.33) / 3 = 2.00\)

- **Few Shot:**
  - First prompting mode: 1.33
  - Every task and every mechanism in one window: 1
  - 11 years old pupil mode: 2.33
  - **Overall Few Shot Average:** \((1.33 + 1 + 2.33) / 3 = 1.55\)

- **Symbolic:**
  - First prompting mode: 1
  - Every task and every mechanism in one window: 1
  - 11 years old pupil mode: 1
  - **Overall Symbolic Average:** \((1 + 1 + 1) / 3 = 1.00\)

- **Arithmetic:**
  - First prompting mode: 1
  - Every task and every mechanism in one window: 2.33
  - 11 years old pupil mode: 1
  - **Overall Arithmetic Average:** \((1 + 2.33 + 1) / 3 = 1.44\)

- **Commonsense:**
  - First prompting mode: 1
  - Every task and every mechanism in one window: 1
  - 11 years old pupil mode: 2.66
  - **Overall Commonsense Average:** \((1 + 1 + 2.66) / 3 = 1.55\)

9. Task A

Zero Shot

First prompting mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

Every mechanism in one window:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate: 3; Average grade: 3

Few Shot

First prompting mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

Every mechanism in one window:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate: 3; Average grade: 3

Symbolic

First prompting mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

Every mechanism in one window:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate: 3; Average grade: 3

Arithmetic

First prompting mode (Chain-of-thought):

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

Every mechanism in one window:

- Accuracy grade: 2; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.33

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate: 3; Average grade: 3

Commonsense

First prompting mode (Chain-of-thought):

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

Every mechanism in one window:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 3; Age-Appropriate: 3; Average grade: 3

Summary

- **Zero Shot:**
  - First prompting mode: 1
  - Every mechanism in one window: 2.33
  - 11 years old pupil mode: 3
  - **Overall Zero Shot Average:** \((1 + 2.33 + 3) / 3 = 2.11\)

- **Few Shot:**
  - First prompting mode: 1
  - Every mechanism in one window: 1
  - 11 years old pupil mode: 3
  - **Overall Few Shot Average:** \((1 + 1 + 3) / 3 = 1.67\)

- **Symbolic:**
  - First prompting mode: 1
  - Every mechanism in one window: 1
  - 11 years old pupil mode: 3
  - **Overall Symbolic Average:** \((1 + 1 + 3) / 3 = 1.67\)

- **Arithmetic:**
  - First prompting mode (Chain-of-thought): 2.33
  - Every mechanism in one window: 2.33
  - 11 years old pupil mode: 3
  - **Overall Arithmetic Average:** \((2.33 + 2.33 + 3) / 3 = 2.56\)

- **Commonsense:**
  - First prompting mode (Chain-of-thought): 1
  - Every mechanism in one window: 1
  - 11 years old pupil mode: 3
  - **Overall Commonsense Average:** \((1 + 1 + 3) / 3 = 1.67\)

9. task B)

Zero Shot

First prompting mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 2; Average grade: 1.33

Every mechanism in one window:

- Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.66

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.66

Few Shot

First prompting mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 2; Average grade: 1.33

Every mechanism in one window:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.66

Symbolic

First prompting mode:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

Every mechanism in one window:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.66

Arithmetic

First prompting mode (Chain-of-thought):

- Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.66

Every mechanism in one window:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 2; Average grade: 1.33

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.66

Commonsense

First prompting mode (Chain-of-thought):

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

Every mechanism in one window:

- Accuracy grade: 1; Alignment with a task: 1; Age-Appropriate: 1; Average grade: 1

11 years old pupil mode:

- Accuracy grade: 3; Alignment with a task: 2; Age-Appropriate: 3; Average grade: 2.66

Summary

- **Zero Shot:**
  - First prompting mode: 1.33
  - Every mechanism in one window: 2.66
  - 11 years old pupil mode: 2.66
  - **Overall Zero Shot Average:** $(1.33 + 2.66 + 2.66) / 3 = 2.22$

- **Few Shot:**
  - First prompting mode: 1.33
  - Every mechanism in one window: 1
  - 11 years old pupil mode: 2.66
  - **Overall Few Shot Average:** $(1.33 + 1 + 2.66) / 3 = 1.66$

- **Symbolic:**
  - First prompting mode: 1
  - Every mechanism in one window: 1
  - 11 years old pupil mode: 2.66
  - **Overall Symbolic Average:** $(1 + 1 + 2.66) / 3 = 1.55$

- **Arithmetic:**

- First prompting mode (Chain-of-thought): 2.66
- Every mechanism in one window: 1.33
- 11 years old pupil mode: 2.66
- **Overall Arithmetic Average:** \((2.66 + 1.33 + 2.66) / 3 = 2.22\)

- **Commonsense:**
- First prompting mode (Chain-of-thought): 1
- Every mechanism in one window: 1
- 11 years old pupil mode: 2.66
- **Overall Commonsense Average:** \((1 + 1 + 2.66) / 3 = 1.55\)

10. task

Zero Shot

First prompting mode:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Every mechanism in one window:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

11 years old pupil mode:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Few Shot

First prompting mode:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Every mechanism in one window:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

11 years old pupil mode:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Symbolic

First prompting mode:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Every mechanism in one window:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

11 years old pupil mode:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Arithmetic

First prompting mode (Chain-of-thought):

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Every mechanism in one window:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

11 years old pupil mode:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Commonsense

First prompting mode (Chain-of-thought):

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Every mechanism in one window:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

11 years old pupil mode:

- Accuracy grade: 4; Alignment with a task: 4; Age-Appropriate: 4; Average grade: 4

Summary

- **Zero Shot:**
  - First prompting mode: 4
  - Every mechanism in one window: 4
  - 11 years old pupil mode: 4
  - **Overall Zero Shot Average:** \((4 + 4 + 4) / 3 = 4\)

- **Few Shot:**
  - First prompting mode: 4
  - Every mechanism in one window: 4
  - 11 years old pupil mode: 4
  - **Overall Few Shot Average:** \((4 + 4 + 4) / 3 = 4\)

- **Symbolic:**
  - First prompting mode: 4
  - Every mechanism in one window: 4
  - 11 years old pupil mode: 4
  - **Overall Symbolic Average:** \((4 + 4 + 4) / 3 = 4\)

- **Arithmetic:**
  - First prompting mode (Chain-of-thought): 4
  - Every mechanism in one window: 4
  - 11 years old pupil mode: 4
  - **Overall Arithmetic Average:** \((4 + 4 + 4) / 3 = 4\)

- **Commonsense:**
  - First prompting mode (Chain-of-thought): 4

- Every mechanism in one window: 4

- 11 years old pupil mode: 4

- **Overall Commonsense Average:** \((4 + 4 + 4) / 3 = 4\)