Investigating Abstraction Capabilities of the o3 Model Using Textual and Visual Modalities

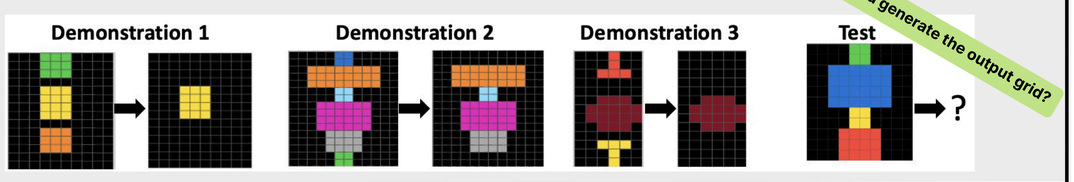
Claas Beger[1], Shuhao Fu[1], Ryan Yi[1], Arseny Moskvichev[2], Melanie Mitchell[1]

[1]Santa Fe Institute, [2]Advanced Micro Devices

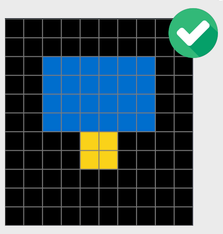
Motivation

OpenAI's o3-preview model outperformed humans on the **Abstraction and Reasoning Corpus**. But does it solve tasks using humanlike abstractions, or less generalizable "shortcuts"?

We investigated this using the **ConceptARC benchmark**, in which simple tasks focus on a single concept (e.g., "top vs bottom").



o3 generates the correct grid...

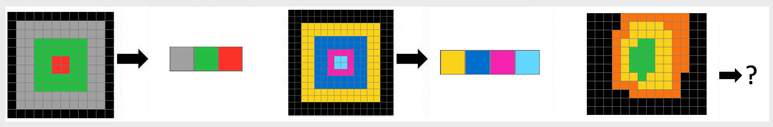


...but it does not use the intended abstraction "top/bottom" and fails on demonstrations 1 and 3!

o3
"Delete every color region that touches any edge of the grid, leaving only regions entirely surrounded by background."

Human
"Copy the grid from the example then remove the topmost and bottom-most shapes. The middle shapes remain unchanged."

How about abstractions that correctly explain the input demonstrations?



o3 generates the correct grid...

...but it does not meet the intended rule, despite solving demonstrations (correct-unintended)!

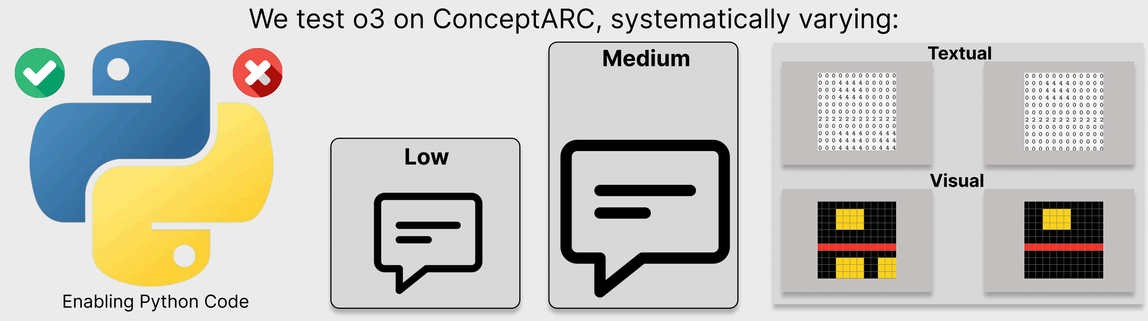
o3
"List all non-background colours in descending order of frequency"

Human
"The grid should be 1 block (height) x width (...). Fill the grid from left to right with the colours, starting from the one used on the outside (...) and finishing with the colour on the inside."

Methodology

Experiments

We test o3 on ConceptARC, systematically varying:



Output

We prompt o3 to generate both **output grid** and **natural-language** rule for each task

Evaluation

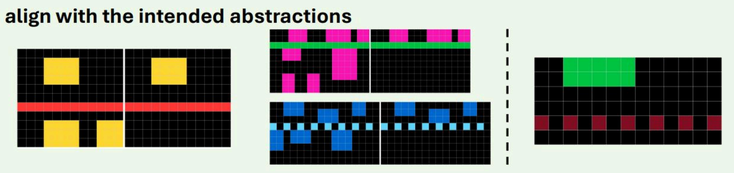
Output-Grid Accuracy: measured via exact match with ground truth

Rule Correctness:

Correct-Intended

Rules that work on the demonstrations and align with the intended abstractions

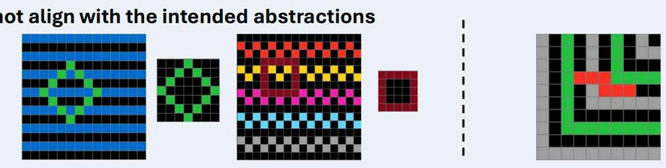
Locate the first row from the top whose colour pattern is identical across the entire width (a solid-colour stripe or a perfectly repeating alternation). Keep that row and everything above it; delete every row beneath it.



Correct-Unintended

Rules that work on the demonstrations, but do not align with the intended abstractions

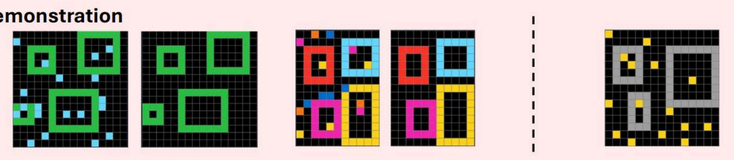
Select the colour that does not appear on the outer border of the input. Cut out the minimal bounding rectangle that contains all cells of this colour. Within that rectangle keep this colour and change every other cell to 0.



Incorrect

Rules that do not work for at least one demonstration

Each cell takes the color that appears most often among its four orthogonal neighbours (up, down, left, right); if no unique majority exists, the cell keeps its original colour.



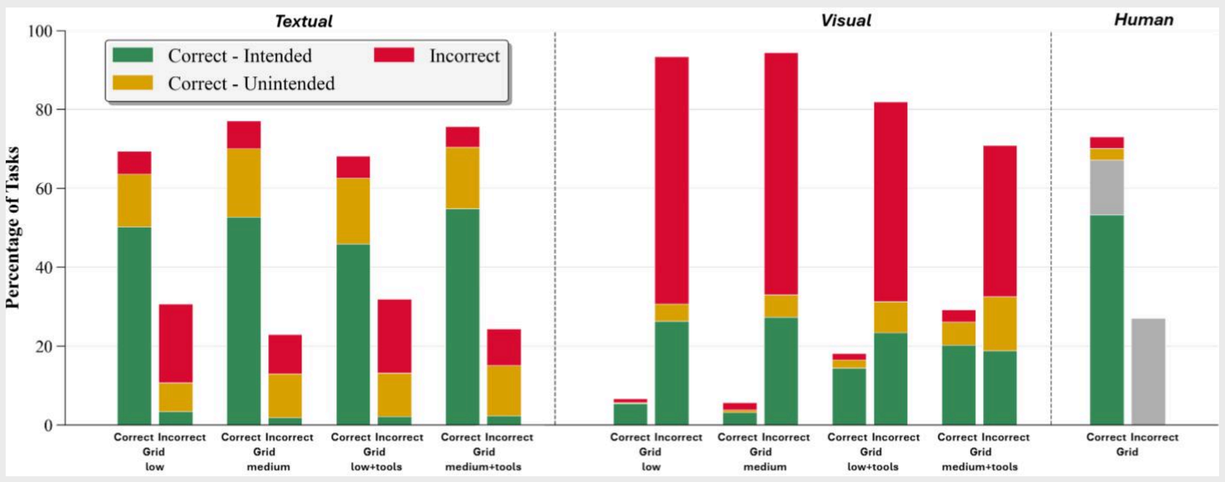
Results

Output-Grid Accuracy

Model	Modality	Low effort	Medium effort	Low effort + tools	Medium effort + tools
o3	Textual	68.3%	**77.1%**	67.9%	75.6%
	Visual	6.7%	5.6%	18.1%	**29.2%**

Textual modality improves with reasoning effort.
Visual modality improves with Python tool use.

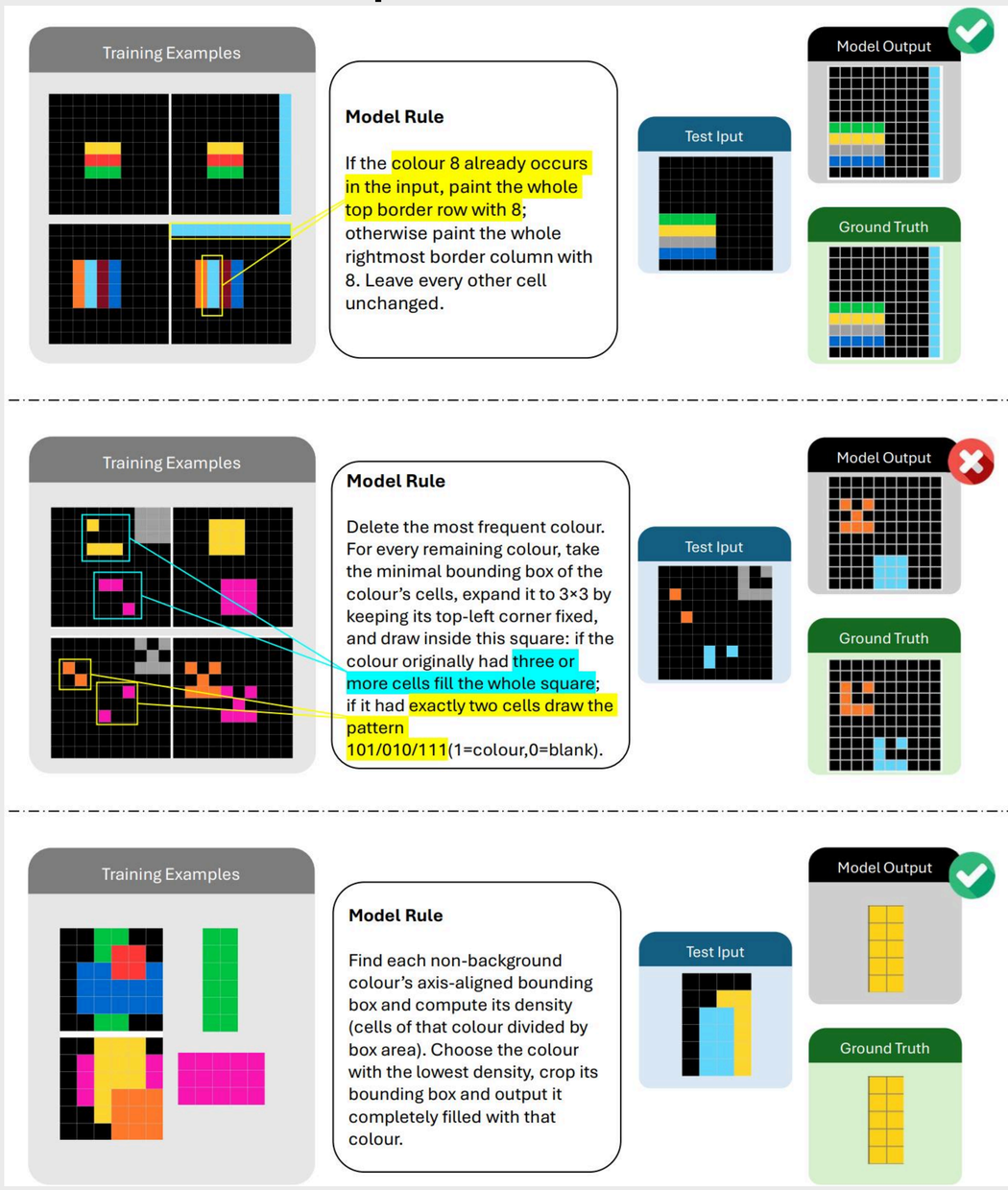
Rule Correctness



For each experimental setting the two bars show the percent of correct and incorrect grid outputs over the 480 ConceptARC tasks. Each bar shows the fraction of tasks for which the rule is *correct-intended*, *correct-unintended*, and *incorrect*. Gray areas in the human-result bars represent rules that we could not classify due to missing data or nonsensical responses.

Textual modality: Accuracy alone overestimates o3's abstract reasoning ability
Visual modality: Accuracy alone underestimates o3's abstract reasoning ability

Examples of o3's "Shortcuts"



Conclusion

Main Takeaway

Evaluating abstract reasoning tasks **solely based on accuracy can be misleading**

- Regardless of modality, **o3 generates correct-intended rules at a lower rate than humans** (around 70% in models vs 90% in Humans)
- o3 frequently produce **shortcuts or heuristics in place of humanlike abstractions**

Textual Modality

While o3 achieved high output-grid accuracy on ConceptARC, it often did so via non-humanlike reasoning. Thus, output accuracy alone **overestimates o3's abstract reasoning abilities** in this modality.

Visual Modality

While o3's output-grid accuracy was very low, it **still produced correct-intended rules fairly often**. Thus, output accuracy alone **underestimates o3's abstract reasoning abilities** in this modality.

Effects of Experimental Variations

- **Increasing reasoning effort** primarily has a positive effect in the **textual** modality
- **Allowing Python-tool usage** mainly helps in **visual** modality