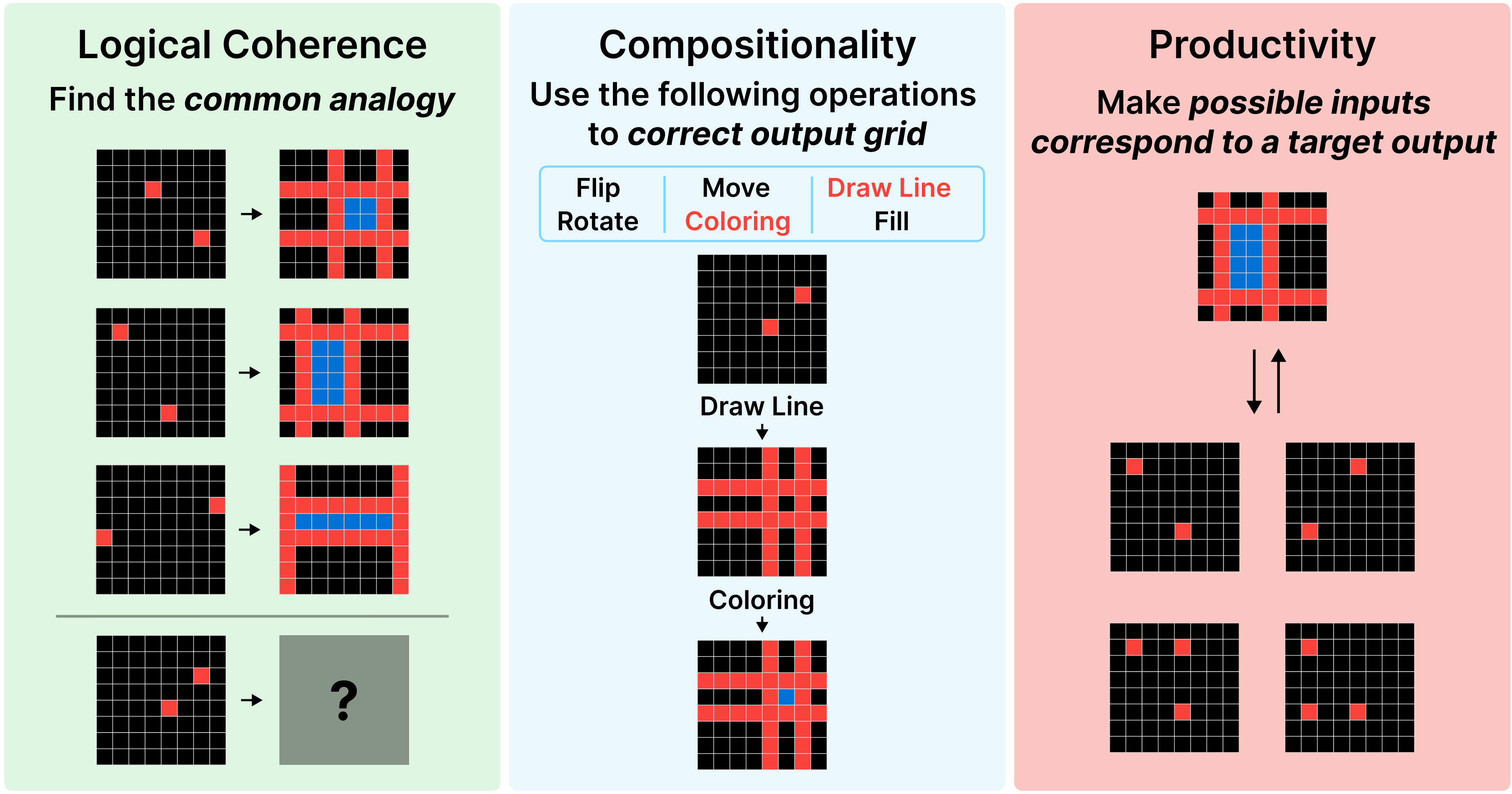
Reasoning Abilities of LLMs: In-Depth Analysis on ARC-AGI

Seungpil Lee* Woochang Sim* Donghyeon Shin* Wongyu Seo Jiwon Park Seokki Lee Sanha Hwang Sejin Kim Sundong Kim‡
Gwangju Institute of Science and Technology

Motivation

To address the lack of systematic reasoning with LLMs, the study adopts the Language of Thought Hypothesis (LoTH) framework, which defines human reasoning through three fundamental capabilities:

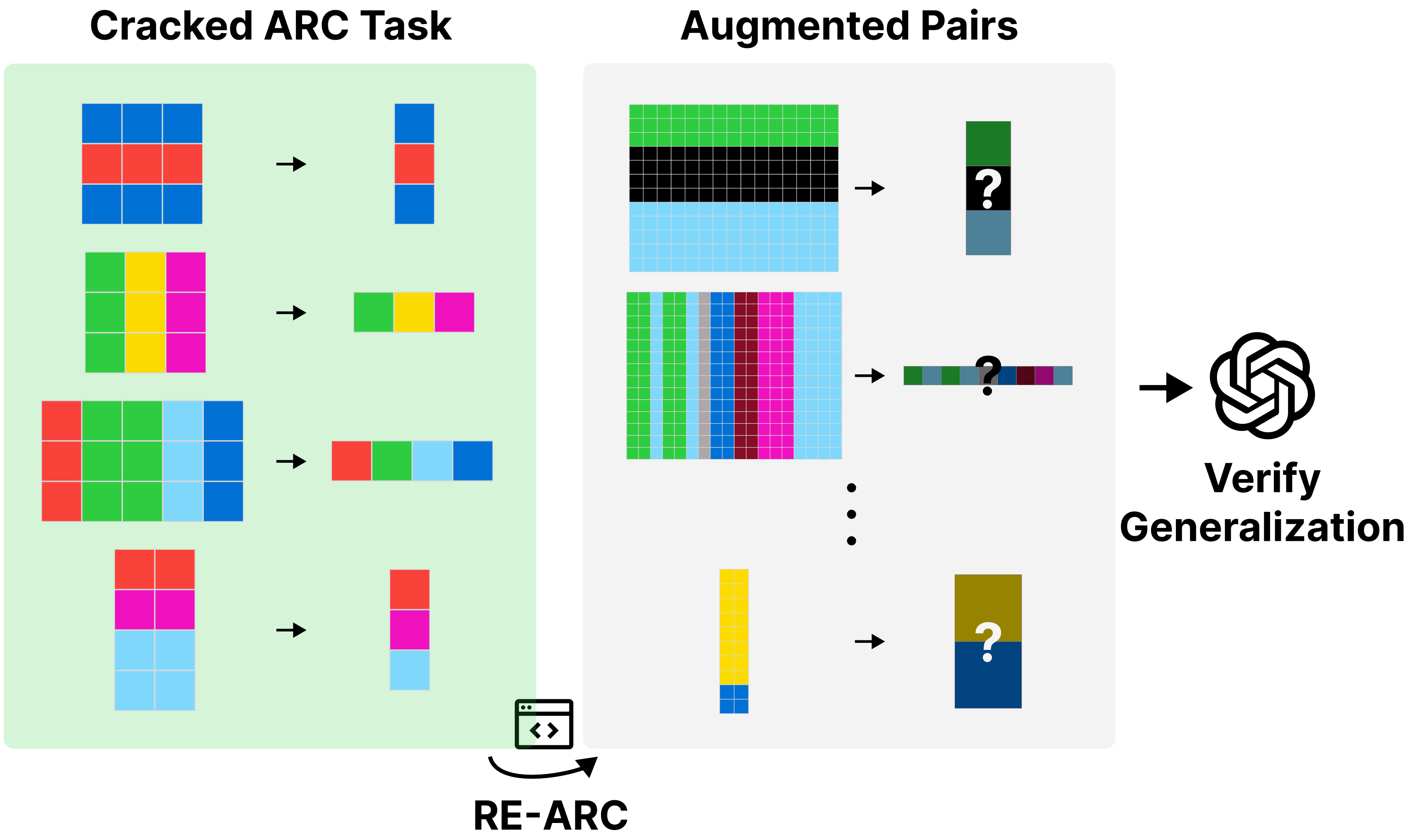
- Logical Coherence: Consistency in reasoning across related tasks.
- Compositionality: Building complex solutions from simple components
- Productivity: Producing infinite solutions from limited components.

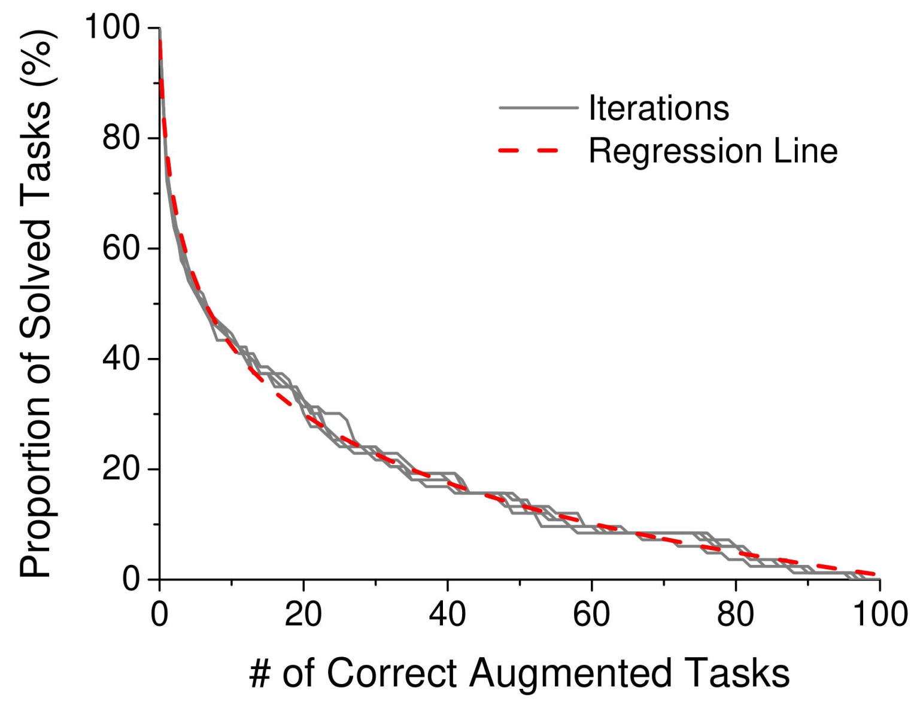


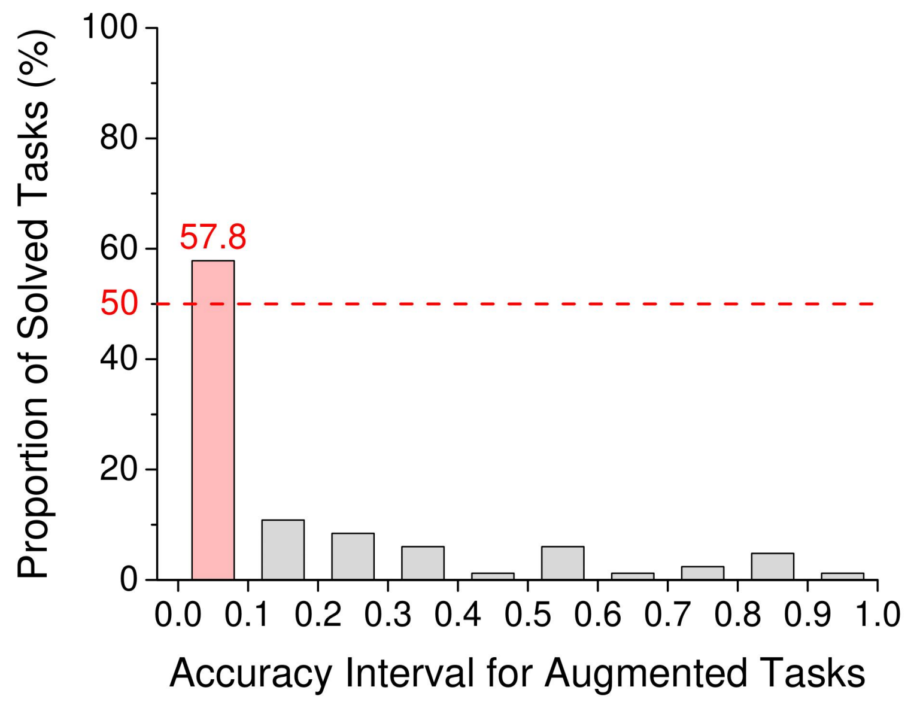
Logical Coherence: Consistency in Reasoning

Logical coherence in the context of this research refers to maintaining consistency in reasoning across tasks, specifically through two dimensions:

- Semantic Coherence: This evaluates how consistently LLMs solve problems based on their reasoning processes and results. In this study, 83 out of 400 tasks were identified as correctly solved by LLMs, which were then analyzed for logical consistency.
- Inferential Coherence: This measures the ability of LLMs to apply logical inferences consistently across similar tasks. Using additional examples generated by a modified ARC benchmark (Re-ARC), the performance of LLMs was tested for their inferential consistency.







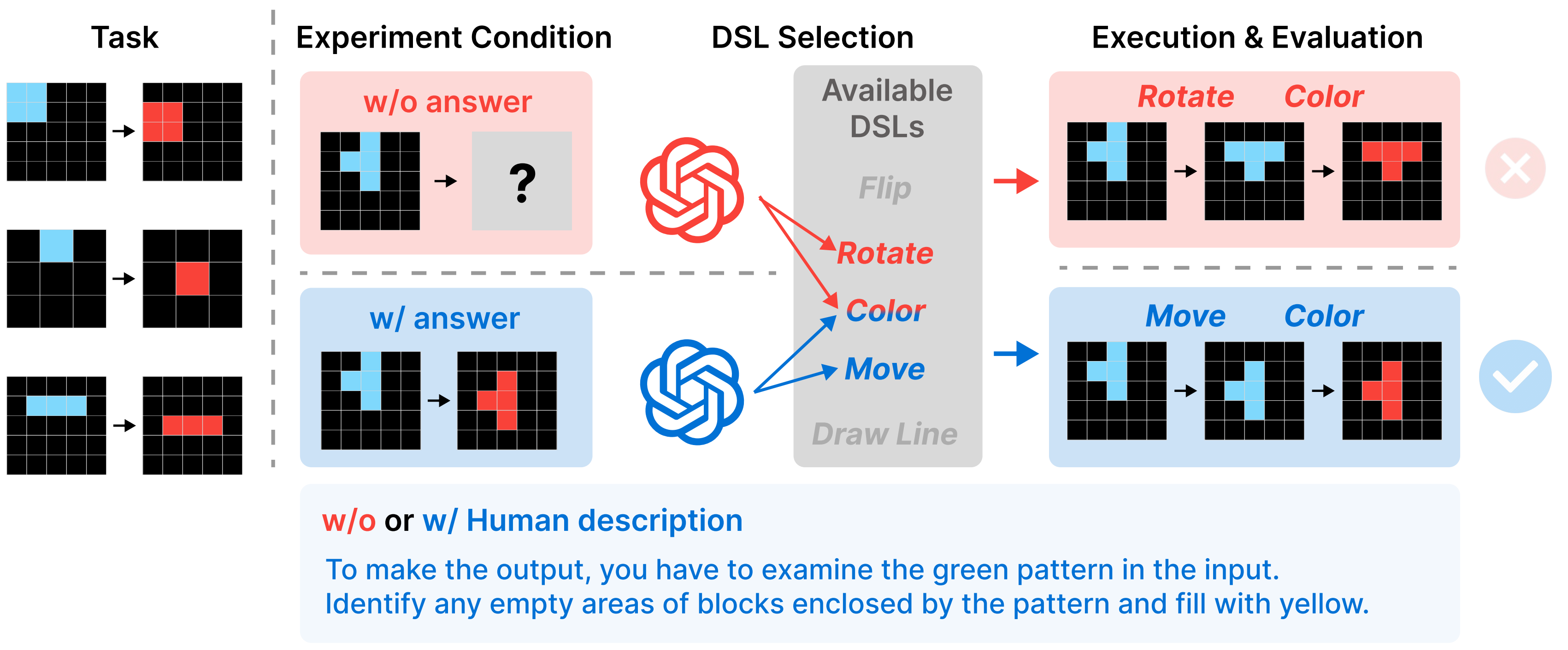
Project page

Paper link

More projects

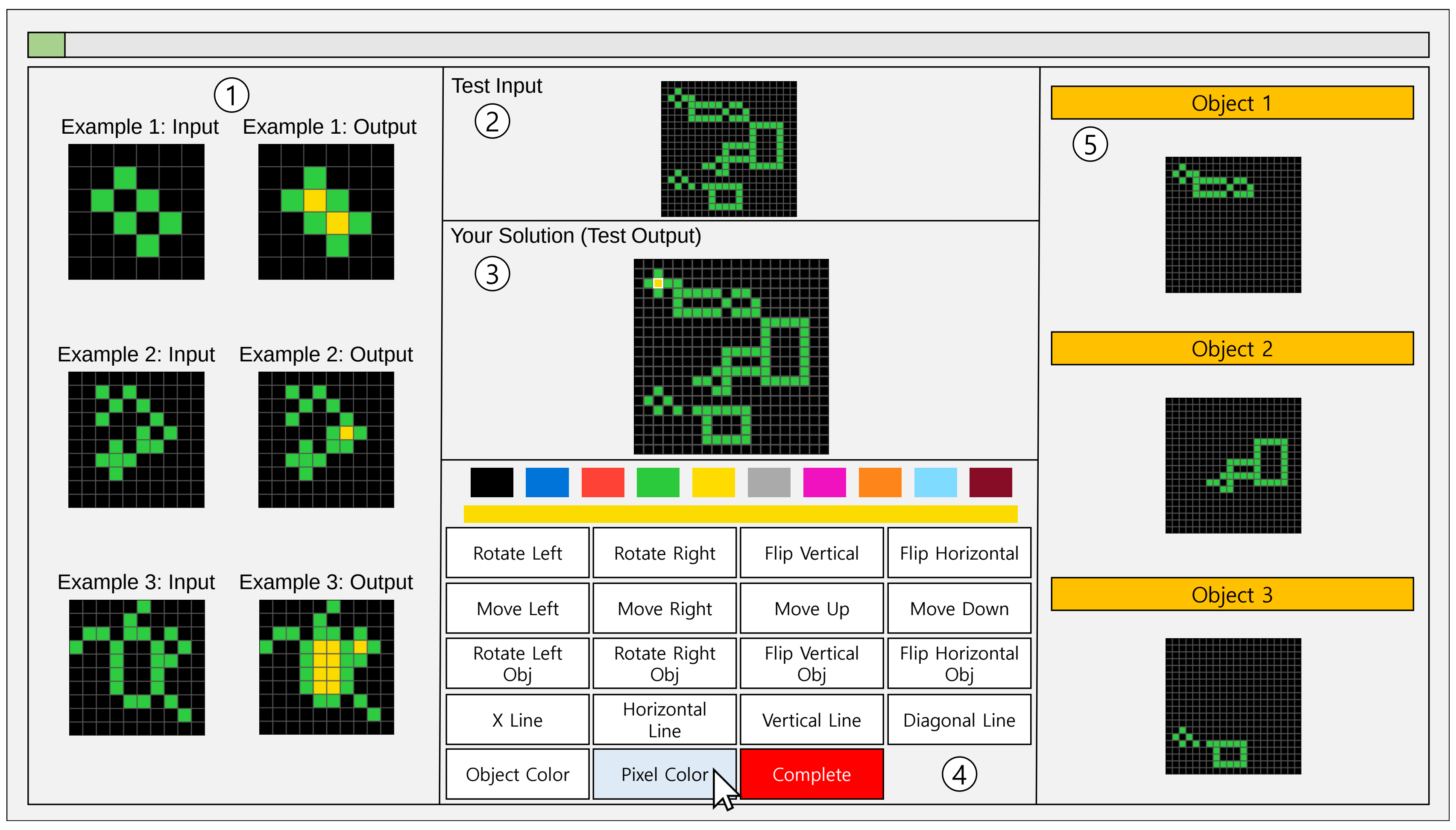
Compositionality: Building Through Components

- Tests LLMs' ability to combine simple DSL functions for ARC task solving
- LLMs achieve 81% in single steps but only 3–14% in complex tasks
- Task accuracy decreases with longer sequences due to cumulative errors

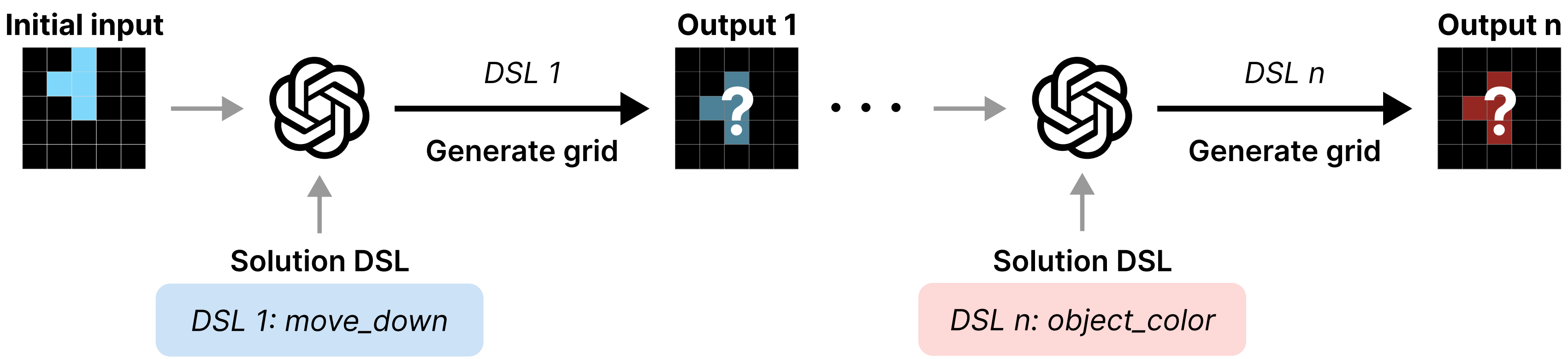


w/o or w/ Human description

To make the output, you have to examine the green pattern in the input.
Identify any empty areas of blocks enclosed by the pattern and fill with yellow.

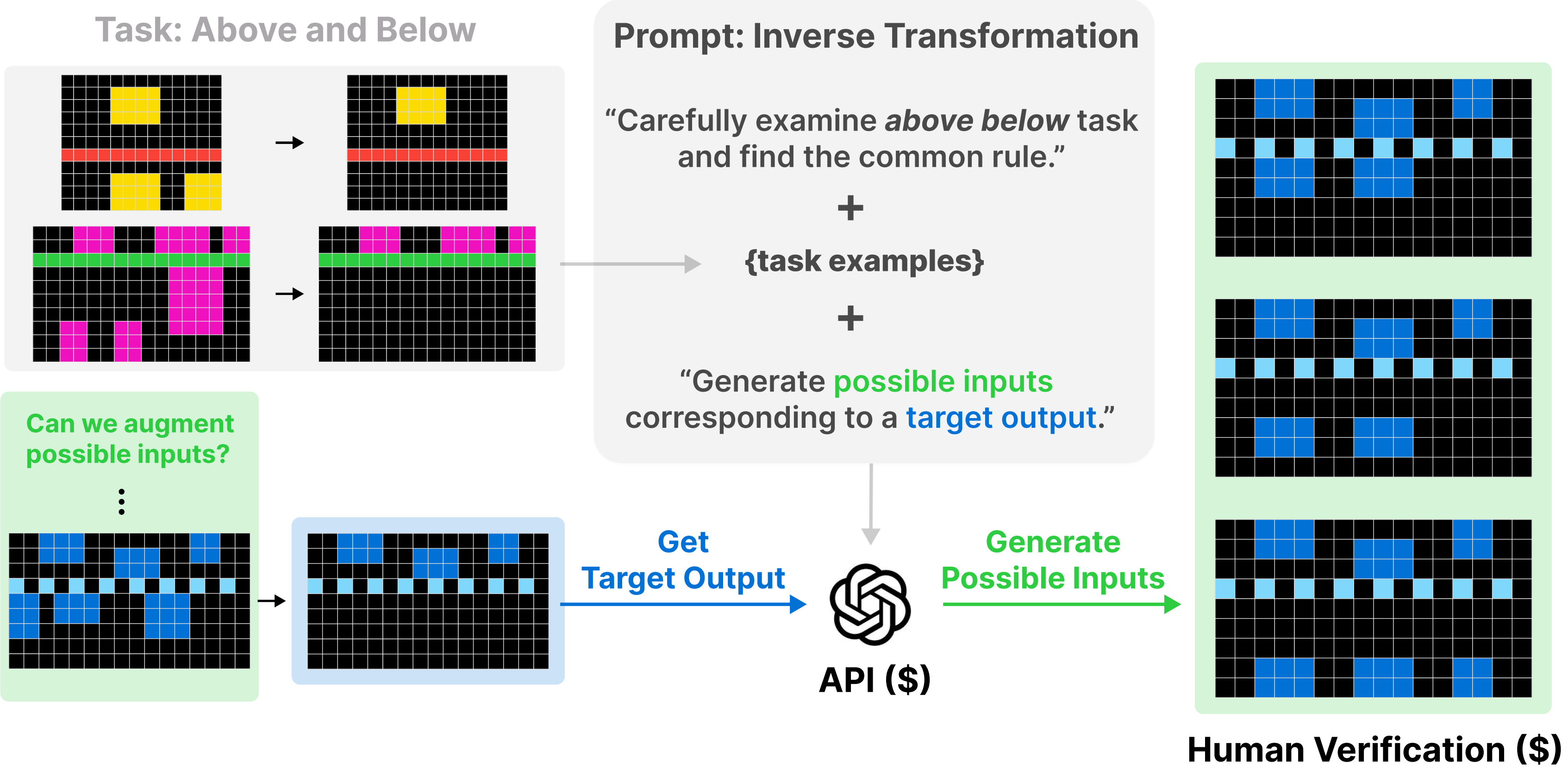


	w/o Human Description	w/ Human Description	Accuracy of Human
w/o Answer	3% (5%)	8% (15%)	86%
w/ Answer	9% (17%)	14% (29%)	X



Productivity: Efficiently Generating Valid Examples

- Testing ability to generate new valid examples from observed patterns
- Cost per valid example (GPT-3.5: $0.0275, GPT-4: $0.3925)
- LLMs mimic patterns rather than understand rules, leading to invalid generations (17.12% valid)



Prompt: Inverse Transformation

"Carefully examine *above below* task and find the common rule."
+
{task examples}
+
"Generate possible inputs corresponding to a target output."

Conclusion

- Logical Coherence: Lack consistency across similar tasks
- Compositionality: Fail in complex tasks (3-14%)
- Productivity: Low valid generation rate (17.12%)