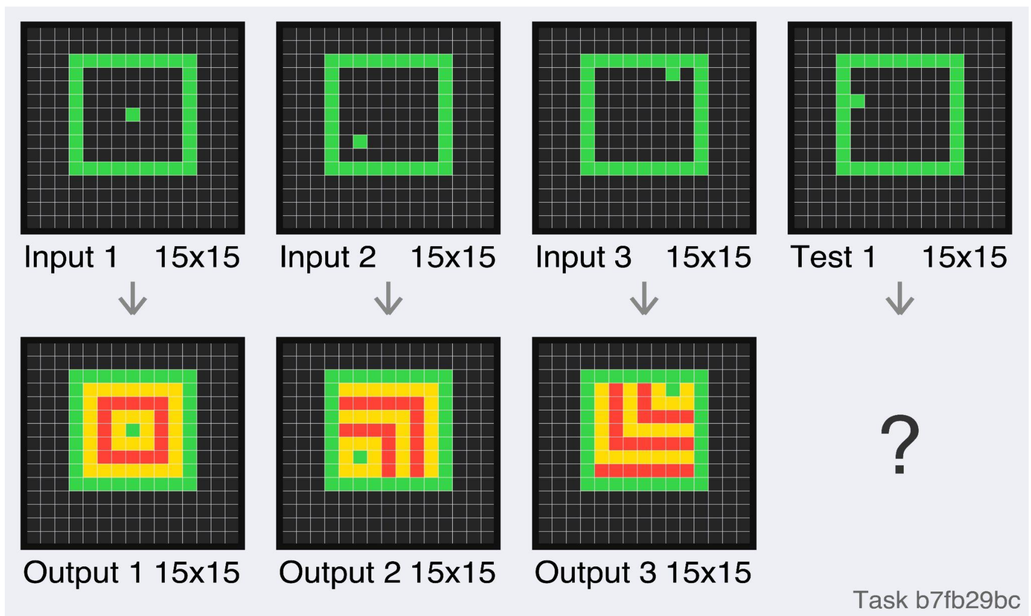
MASR: Multi-Agent System with Reflection for the Abstraction and Reasoning Corpus

Kiril Bikov, Mikel Bober-Irizar, Soumya Banerjee

UNIVERSITY OF CAMBRIDGE

Abstraction and Reasoning Corpus (ARC)

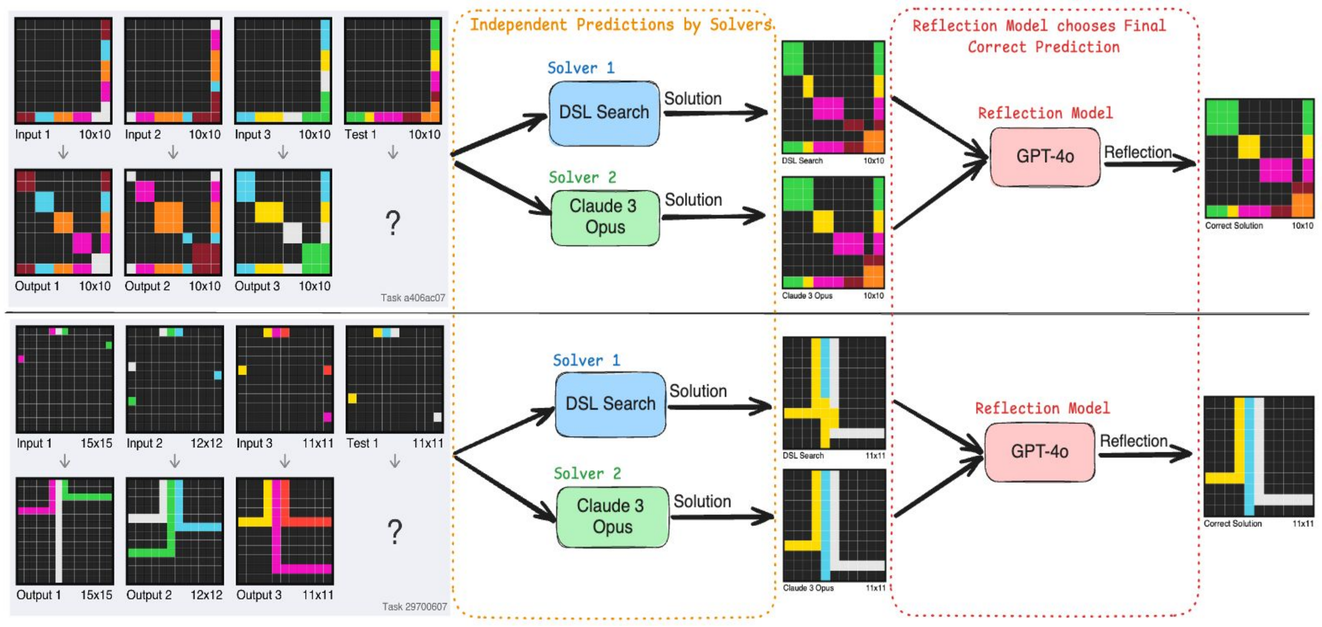
ARC addresses the gap between human intelligence and AI models. IT consists of 1000 visual tasks, capturing essential aspects of abstraction and analogy. Previous solvers of ARC have either been single LLMs or heuristic search systems. We explore how different systems can be combined in a multi-agent setting on ARC.



MASR: Multi-Agent System with Reflection

MASR relies on agents of various architectures: LLMs or Domain Specific Languages using Program Synthesis. When predicting the correct solution to an ARC task, MASR executes in two main stages:

1. **Independent predictions by each agent.**
2. **Reflection over predictions to choose final one.**



AugARC: Augmented ARC for LLMs

To tackle the **limited number of ARC training tasks**, we propose the following augmentation techniques:

- *Rotation*: of each ARC grid by 90° or 270°.
- *Flipping*: *horizontally and vertically.*
- *Permutations*: rearranges input-output pairs.

The AugARC sets vary from 2000 to 18 million tasks.

3-Shot AugARC Benchmark - a unified, easy way to evaluate LLMs on reasoning. Each ARC task starts with a textual description, the ARC grids are represented as a 2D matrix of numbers.

Flexibility of MASR

MASR allows any number of agents to be used, as the reflection model can easily process the outputs of various agents. This makes MASR a highly flexible and customisable architecture, as each of its components - the agents and the reflection model, can easily be changed.

Results of base LLMs on ARC and AugARC

Model	ARC	AugARC	Increase
Llama-2 7B	5/400	7/400	29%
Mistral 7B	9/400	15/400	67%
Llama-2 13B	5/400	8/400	100%
Llama-2 70B	7/400	14/400	100%
Mixtral 8x7B	9/400	18/400	**125%**
Gemini Pro	20/400	**33/400**	65%

Performance of MASR Configurations

Solver 1	Solver 2	Solver 3	Reflection Model	ARC Correct
DSL Search	Claude 3 Opus	-	Llama-3 70B	133/400
DSL Search	Claude 3 Opus	-	GPT-4-turbo	165/400
DSL Search	Claude 3 Opus	-	GPT-4o	**166/400**
DSL Search	Claude 3 Opus	Fine-Tuned Llama-3 8B	Claude 3.5 Sonnet	163/400

Performance of Fine-tuned LLMs on AugARC

Model	Base	Fine-tuned	Increase
Llama-2 7B	7/400	21/400	**200%**
Mistral 7B	15/400	23/400	53%
Llama-2 13B	8/400	18/400	125%
Llama-3 8B	21/400	**34/400**	62%

MASR against Previous Approaches

Program Synthesis	Brute Force	26/400
	Neurodiversity solver	45/400
	DSL Search	160/400
Ensemble	Voting	161/400
Multi-agent	**MASR**	**166/400**

https://github.com/kiril-bikov/AugARC