Bridging the ARC from Intrigue to Innovation

OR

Would you quit your PhD for $600,000?

Jack Bell, PhD candidate & Researcher
Department of Computer Science
14.02.25

Outline

- Background & Historical Context
- Defining the ARC Prize & Benchmark
- Approaches & Strategies
- Open Questions & Future Directions
- Conclusion & Q&A

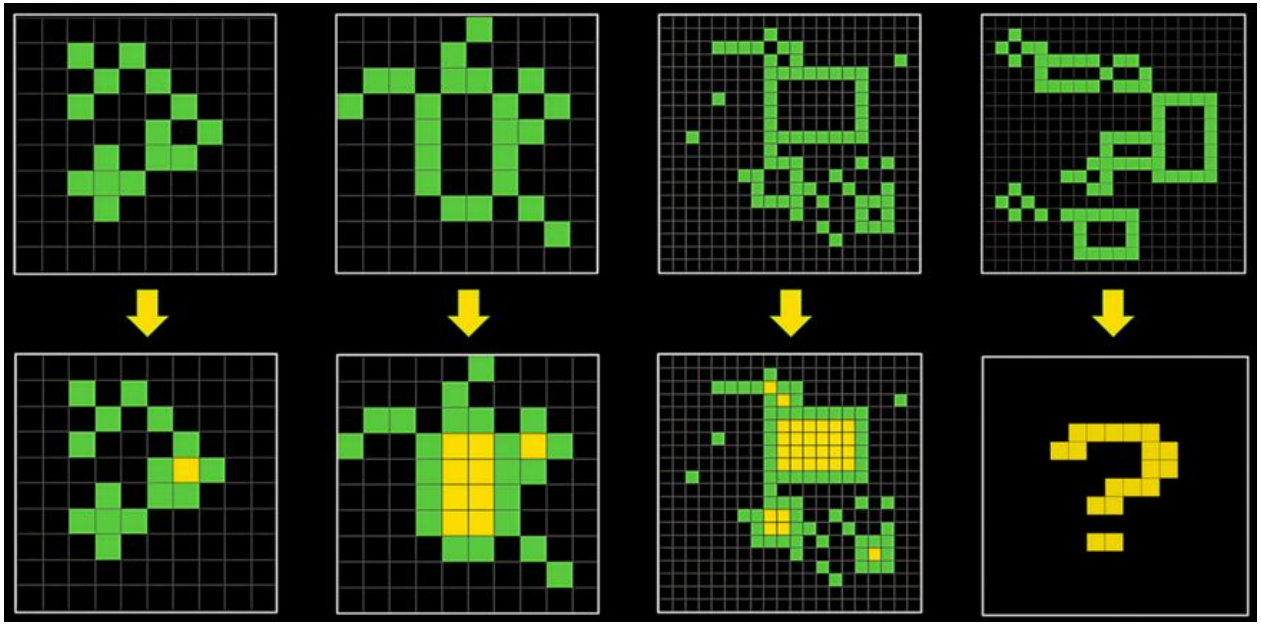
The Origins of the ARC Prize

- Abstract and Reasoning Corpus for Artificial General Intelligence (ARC-AGI) released in 2019
- After 2 “ARCathons” in 2022 and 2023, the ARC Prize was released in 2024 with a $1.1 million prize pool - with $600,000 for reaching 85% accuracy
- Intended as a “stepping stone” in the direction of AGI

Informally speaking, **AGI** is a system that can efficiently acquire new skills outside of its training data.

Benchmark Design

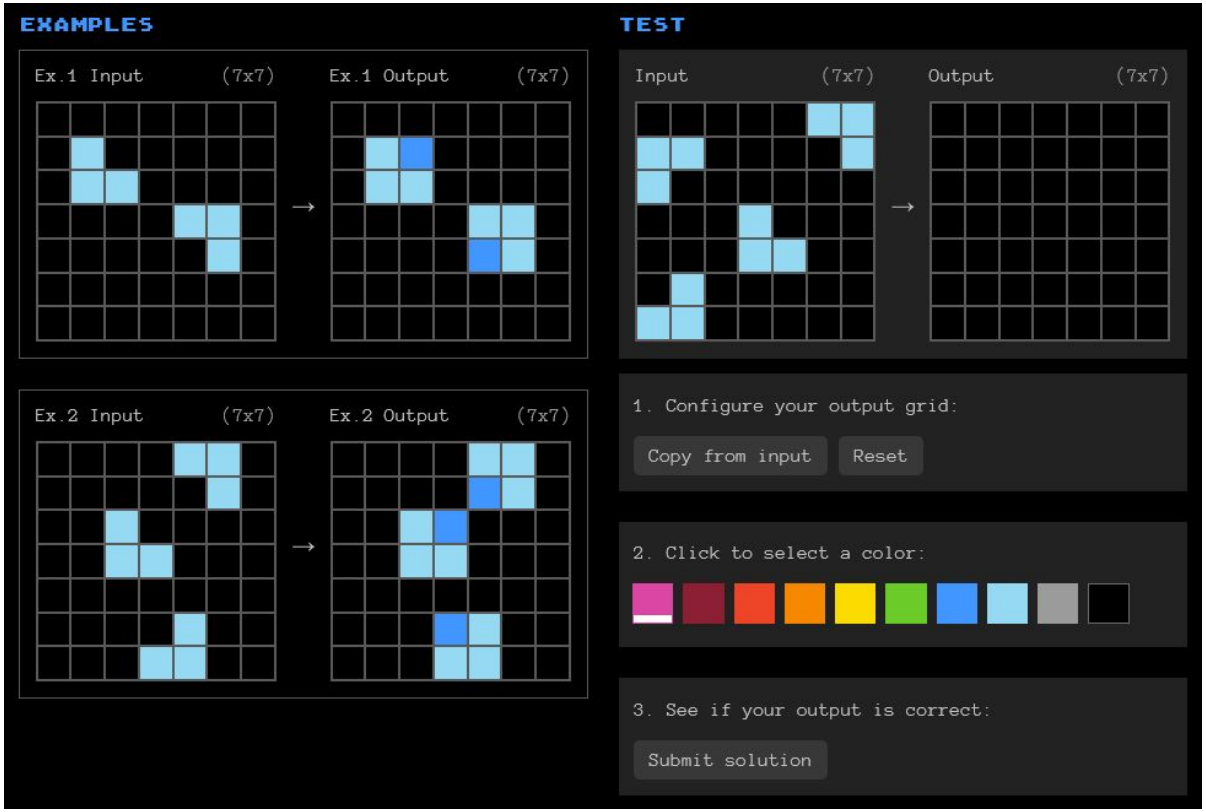
- Input, output examples (variable size from 1x1 to 30x30)
- Each square can be one of ten colours
- Pixel perfect output with correct dimensions
- Resistant to memorisation



https://github.com/fchollet/ARC-AGI

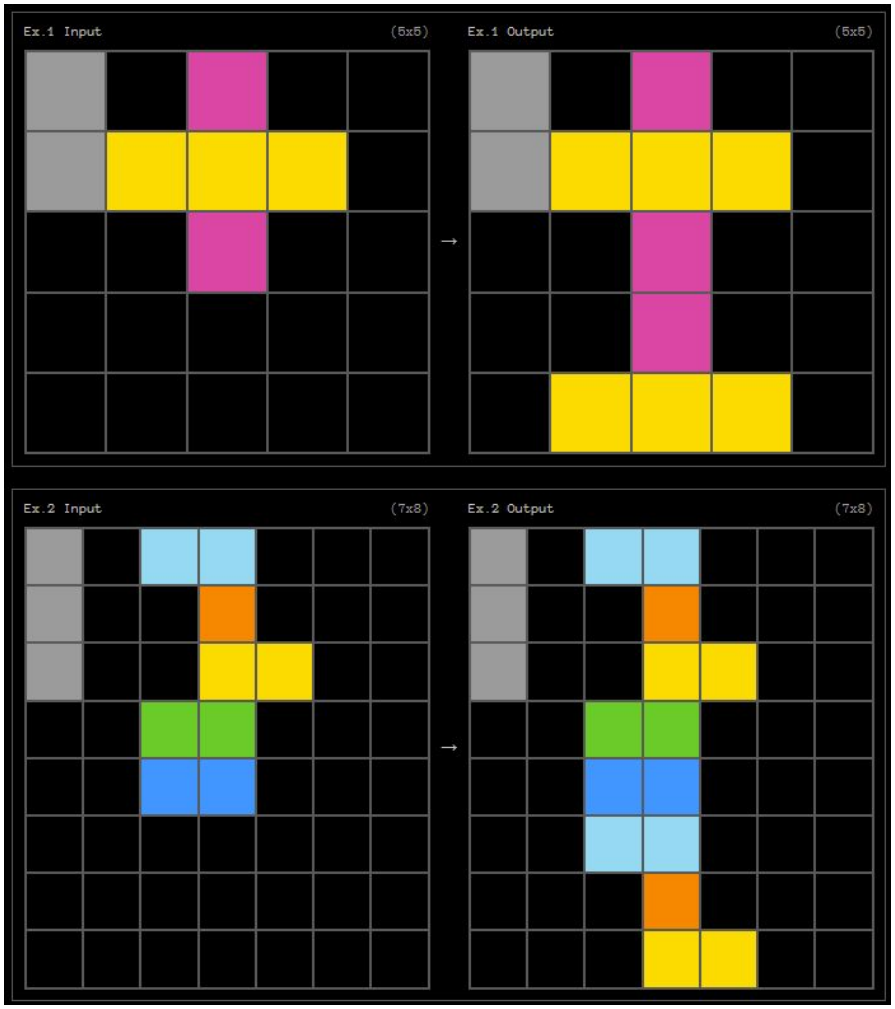
Benchmark Design

Train | Test | Eval - 400 | 400 | 100



https://arcprize.org/

Benchmark Design



https://arcprize.org/

Assumed Knowledge

Objectness - objects can interact, but can't appear or disappear without reason.

Goal-directedness - objects can pursue goals.

Numbers & Counting - objects can be counted or sorted by shape, appearance or movement.

Basic Geometry & Topology - objects can have basic shapes and can be manipulated by transformations like rotations.

Chollet (2019)

Setup and Constraints

Private Leaderboard - linked to prize money

Limited Resources - in notebook format, 12 hours runtime of compute (4vCPUs) or GPU (one P100 or two T4) ~$10

No Internet

Private evaluation set

Public Leaderboard

$10,000 of compute on Kaggle

Internet access

'**Semi-private**' evaluation set

Scoring and Human Performance

2 submissions per task - the best submission counts
(Semi-)private evaluation set - 100 hard tasks with pixel perfect grid required

Aim: Reach **85%** accuracy on private test set

Average human: **64.2%** on public eval set (LeGris *et al*., 2024)

Max score of 10 “high-IQ” individuals: **~97%**

99% of public eval tasks solved by one Mechanical Turk worker - with 10 workers assigned to each task (Chollet, 2025)

Potential Impact

Reduce reliance on huge datasets

- Few-shot adaptation should improve data-efficiency

Towards generalisable models

- Better model architectures may be required to efficiently solve ARC-AGI

Unlocking program synthesis

- Input output pair task generalisation will be possible in future enabling programming for all

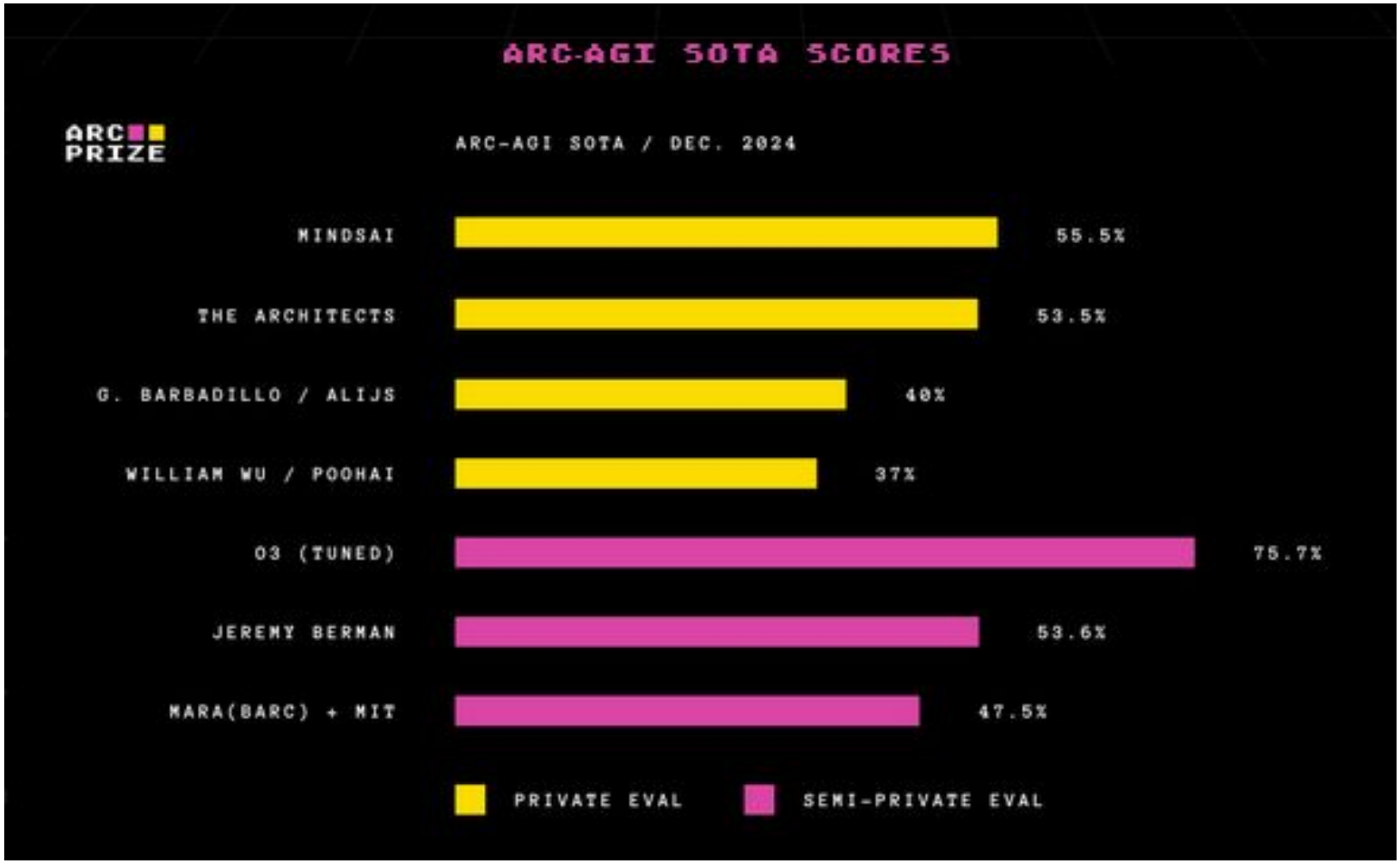
High-level Approaches

- Domain Specific Language (DSL) - brute force search, 40% *private* (Liukis, 2024)
- Test Time Training (TTT) - 47.5% *semi-private,* (Akyürek *et al*., 2024)
- Program Synthesis - 53.6% *semi-private* (Berman, 2024)
- Ensemble Solutions - ***theoretically*** up to 80% - but using a lot of compute!
- Ensemble of brute force search techniques (~50%)
- Direct LLM prompting - **best:** 10% accuracy, original GPT-3: 0%

Why do LLMs underperform?

- LLMs are incredibly **skilled** at specific tasks they have enough training examples for
- But, they are simply doing **pattern recognition**. Without a template to solve the *exact* problem, they can struggle to solve a Caesar cipher (McCoy *et al*., 2023) other than for the three most common shifts (13, 3, 1)
- Without general **reasoning** abilities (e.g. System 2 thinking in humans), solving **novel** tasks through learning new concepts is difficult since they are **static**

State of the art - is compute the limit?



https://arcprize.org/

Program Synthesis

LLM-powered program generation in open-ended languages
Train an LLM on programming related data and try to create thousands of Python program which could solve the task. (Berman, 2024)

LLM-powered iterative program debugging
Use LLMs to iteratively debug nearly-correct programs as selected by a heuristic (Greenblatt, 2024)

Test Time Training (TTT)

Idea: Solve the task by creating a different version of the LLM fine-tuned exactly to the problem you are trying to solve

Fine-tuning a pretrained LLM at test time
Create a fine-tuned LLM for each task

Leverage data augmentation
Create many examples using datasets like ReARC (Hodel, 2024) which are semantically similar to the original task

Evolutionary Test Time Compute

Initial Generation
Generate a number of python programs to try and solve the task by induction

Fitness Evaluation
Use puzzle correctness or pixel accuracy on the test examples to decide on fitness

Selection & Reproduction
Select best prompts as parents and create revisions and offspring

Iterate until solution found or max generation limit is hit

(Berman, 2024)

Challenges and Complexities

Resistant to ‘memorisation’
Requires generalisation from ‘Core’ knowledge to perform well on eval set

Private eval score is shown on solution upload
You can infer something about the private test set by iteratively uploading your solution

LLMs pretrained on Github data
Since ARC-AGI is hosted on github and publicly available, LLMs will be pretrained on this data

Could be brute-forced
With enough hand-crafted examples, the solution could be learned

Future Directions

Deep Learning Guided Program Synthesis
Train a deep learning guided program synthesis search, possibly using an LLM to guide the search

Continual Learning
Is there a way to leverage test-time *adaptation* to leverage the similarity between different test tasks? (Sun *et al.*, 2025)

ARC-AGI 2
A new benchmark is being released this year to try and address the shortcomings of the first dataset (Chollet and Knoop, 2024)

$600,000 dollar prize for reaching 85% with $10 of compute and no internet access

If you would like to collaborate, please reach out!

References (1)

- Arc prize ARC Prize. Available at: https://arcprize.org/ (Accessed: 12 February 2025).
- Francois Chollet ARC-AGI: The abstraction and reasoning corpus, GitHub. Available at: https://github.com/fchollet/ARC-AGI (Accessed: 12 February 2025).
- Chollet, F., 2019. On the measure of intelligence. arXiv preprint arXiv:1911.01547.
- LeGris, S., Vong, W.K., Lake, B.M. and Gureckis, T.M., 2024. H-ARC: A Robust Estimate of Human Performance on the Abstraction and Reasoning Corpus Benchmark. arXiv preprint arXiv:2409.01374.
- Liukis, A. (2024) ARC prize 2024, Kaggle. Available at: https://www.kaggle.com/code/gregkamradt/arc-prize-2024-solution-4th-place-score-40-811b72 (Accessed: 12 February 2025).
- Akyürek, E., Damani, M., Qiu, L., Guo, H., Kim, Y. and Andreas, J., 2024. The surprising effectiveness of test-time training for abstract reasoning. arXiv preprint arXiv:2411.07279.
- Berman, J. (2024) How I came in first on Arc-AGI-pub using sonnet 3.5 with evolutionary test-time compute. Available at: https://jeremyberman.substack.com/p/how-i-got-a-record-536-on-arc-agi (Accessed: 12 February 2025).

References (II)

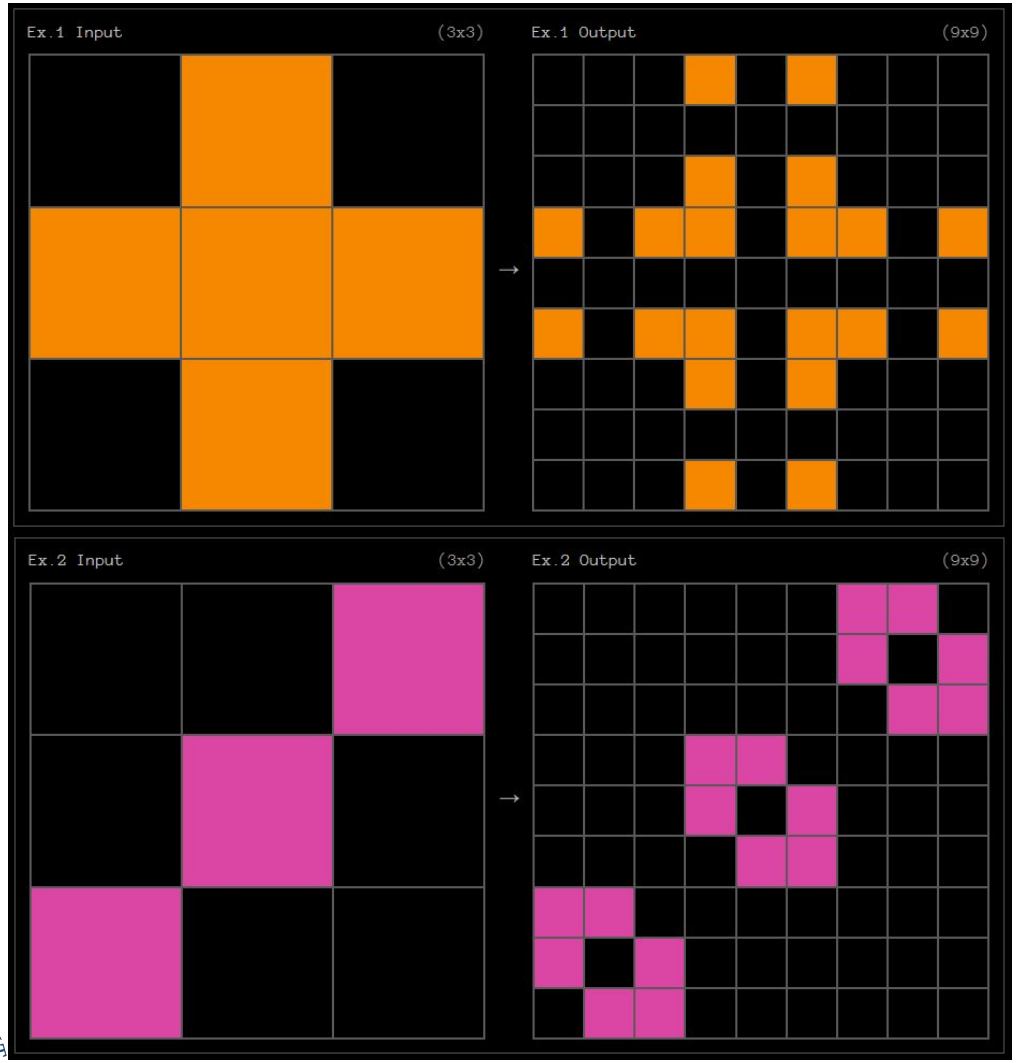
- Greenblatt, R. (2024) Getting 50% (SOTA) on ARC-AGI with gpt-4o. Available at: https://redwoodresearch.substack.com/p/getting-50-sota-on-arc-agi-with-gpt (Accessed: 12 February 2025).
- Hodel M. (2024) RE-ARC: Reverse engineering the abstraction and Reasoning Corpus. Available at: https://github.com/michaelhodel/re-arc (Accessed: 12 February 2025).
- Sun, Q., Cetin, E. and Tang, Y., 2025. $\text {Transformer}^ 2$: Self-adaptive LLMs. arXiv preprint arXiv:2501.06252.
- Chollet F. and Knoop M. (2024) Francois Chollet, Mike Knoop-LLMs won't lead to AGI-$1,000,000 Prize to find true solution. [Podcast]. 21 June 2024. Available at: https://www.youtube.com/watch?v=UakqL6Pj9xo (Accessed 12 February 2025)
- Chollet F. (2025) Francois Chollet-ARC reflections-NeurIPS 2024. [Podcast]. 9 January 2025. Available at: https://open.spotify.com/episode/087N01jW5MUNdoH9AKEXYB (Accessed 12 February 2025)
- McCoy, R.T., Yao, S., Friedman, D., Hardy, M.D. and Griffiths, T.L., 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. Proceedings of the National Academy of Sciences, 121(41), p.e2322420121.

Q&A / Discussion

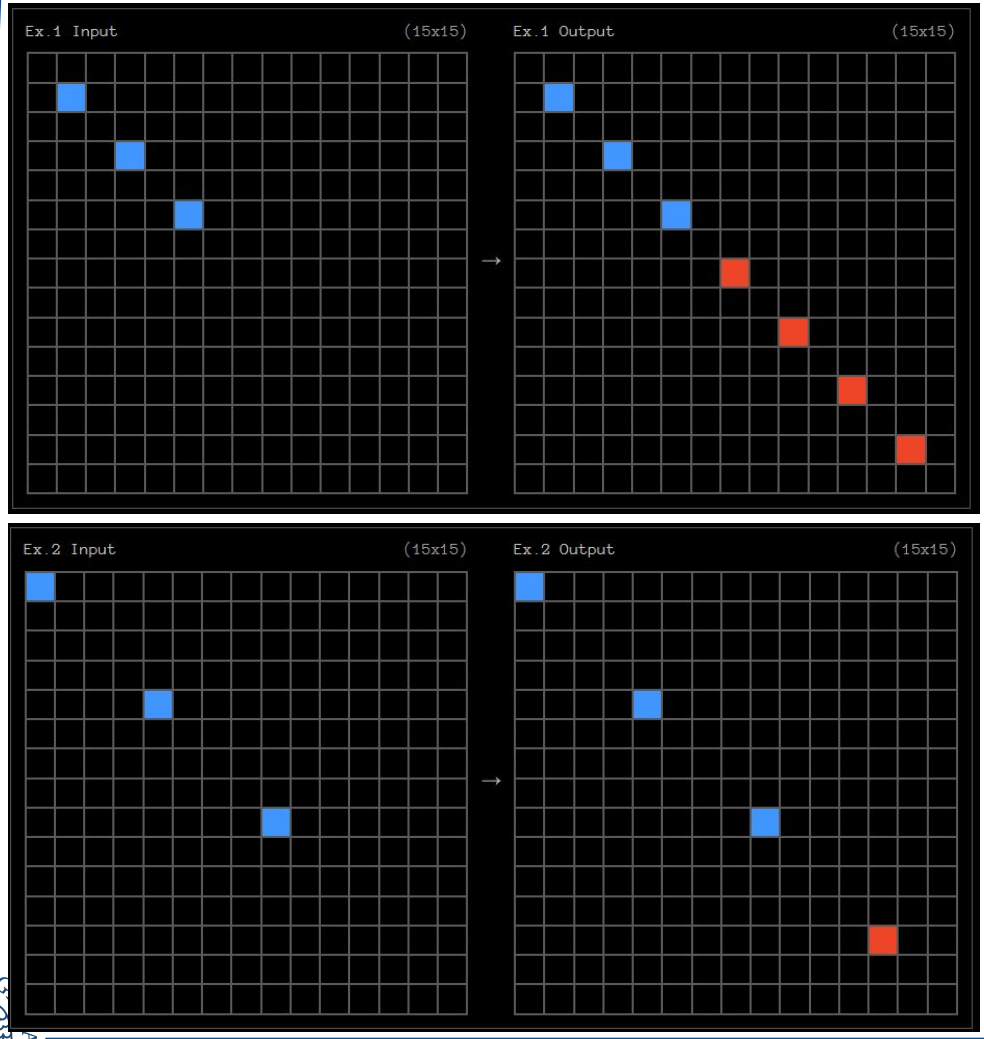
If you have ideas of how to progress towards the ARC prize, I'd love to hear them!

Any Questions (or requests to see more puzzles)?

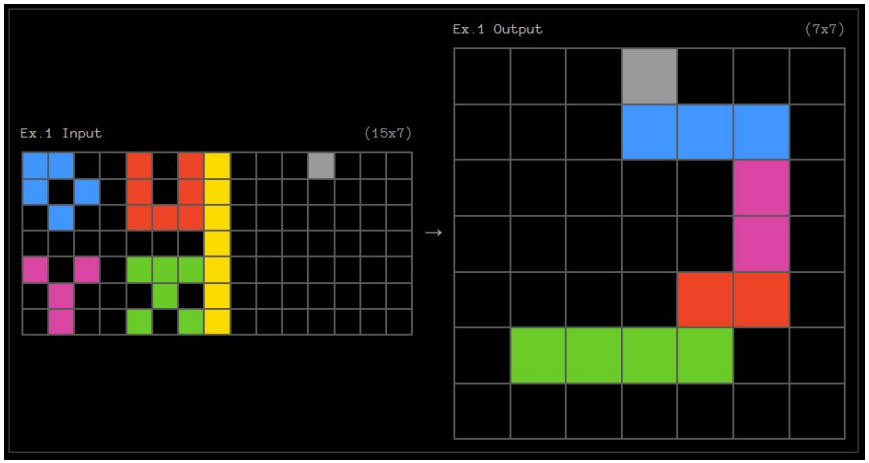
Extra Problems (I)

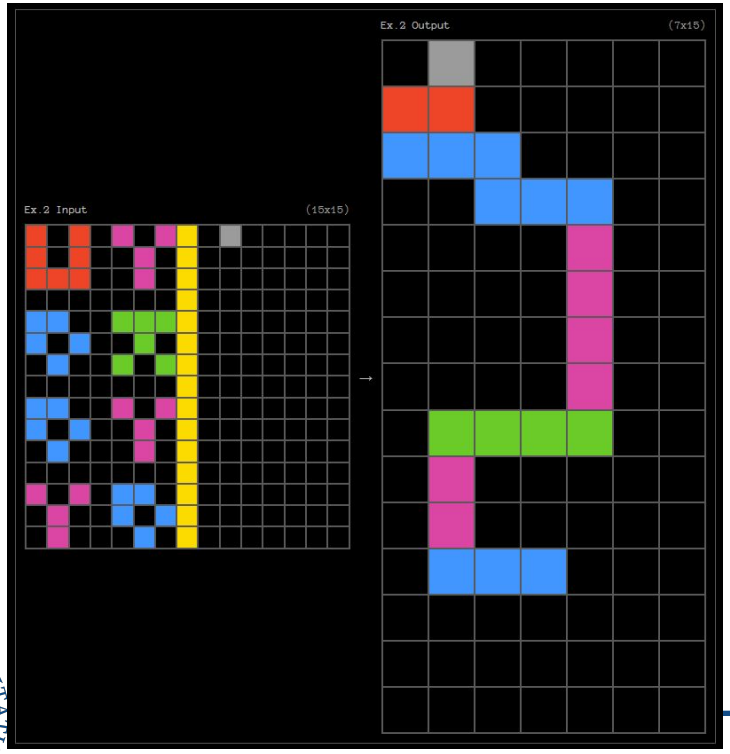


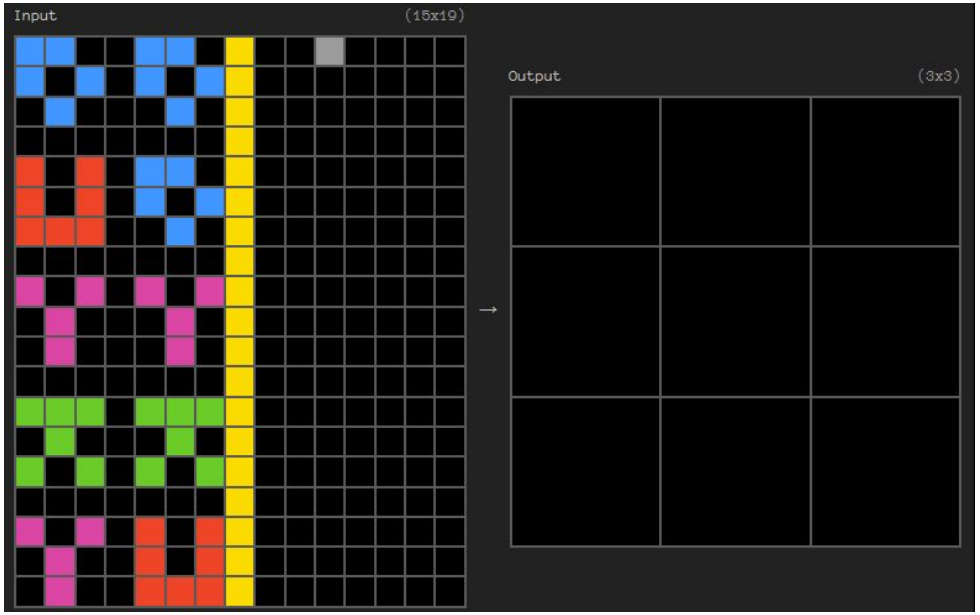
Extra Problems (II)



Extra Problems (III)







Extra Problems (IV)