COMMUNICATING NATURAL PROGRAMS TO HUMANS AND MACHINES

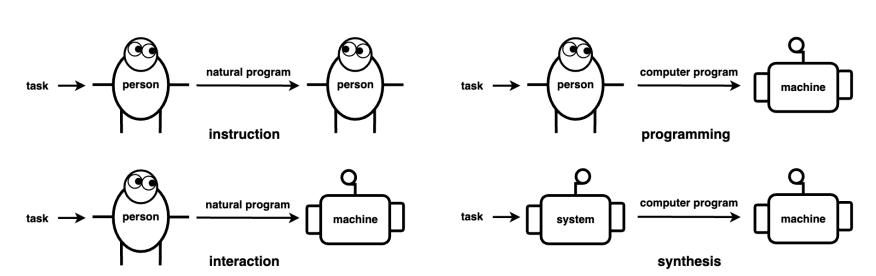
Sam Acquaviva¹, Yewen Pu ², Marta Kryven¹, Theodoros Sechopoulos¹, MH Tessler¹, Josh Tenenbaum

¹Massachusetts Institute of Technology {samacqua, mkryven, theosech, tessler, jbt}@mit.edu ² yewen.pu@autodesk.com



How to build generalizable AI?

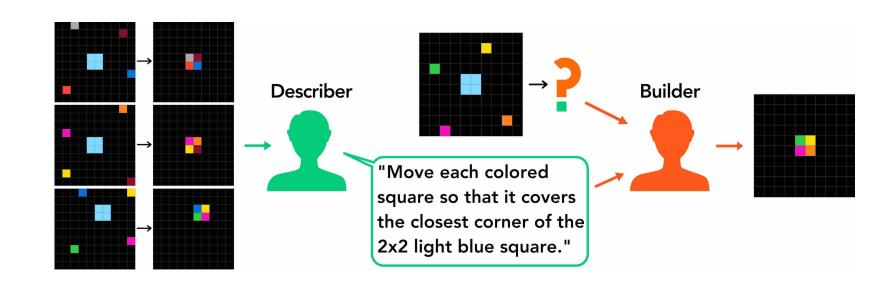
Goal: Build a program synthesizer augmented with natural language to solve program-induction problems in a human-like way



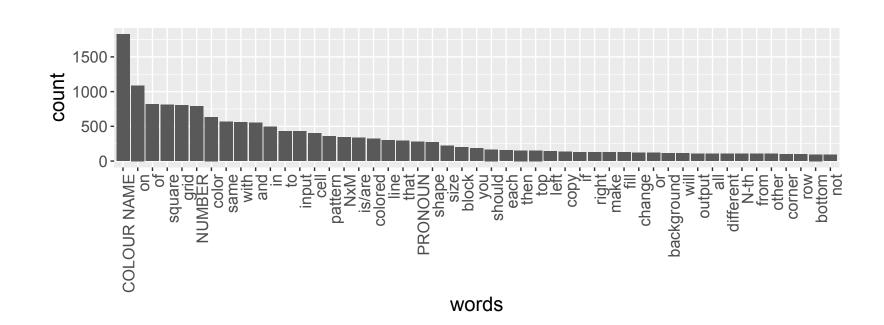
Four ways humans and machines communicate with each other: instruction (top-left), programming (top-right), interaction (bot-left), and synthesis (bot-right).

The Abstraction and Reasoning Corpus (ARC) a set of program induction tasks to evaluate general intelligence of an AI[1]. Designed to leverage Core Knowledge[2] For humans 80% of ARC tasks are trivial[3]. Best Kaggle contestant solved 20%[4].

Method: Collect Language-augmented ARC (LARC) dataset: "natural programs" – instructions given by humans to describe how to solve ARC.



The Bandit Game procedure used to collect natural language annotations for LARC.



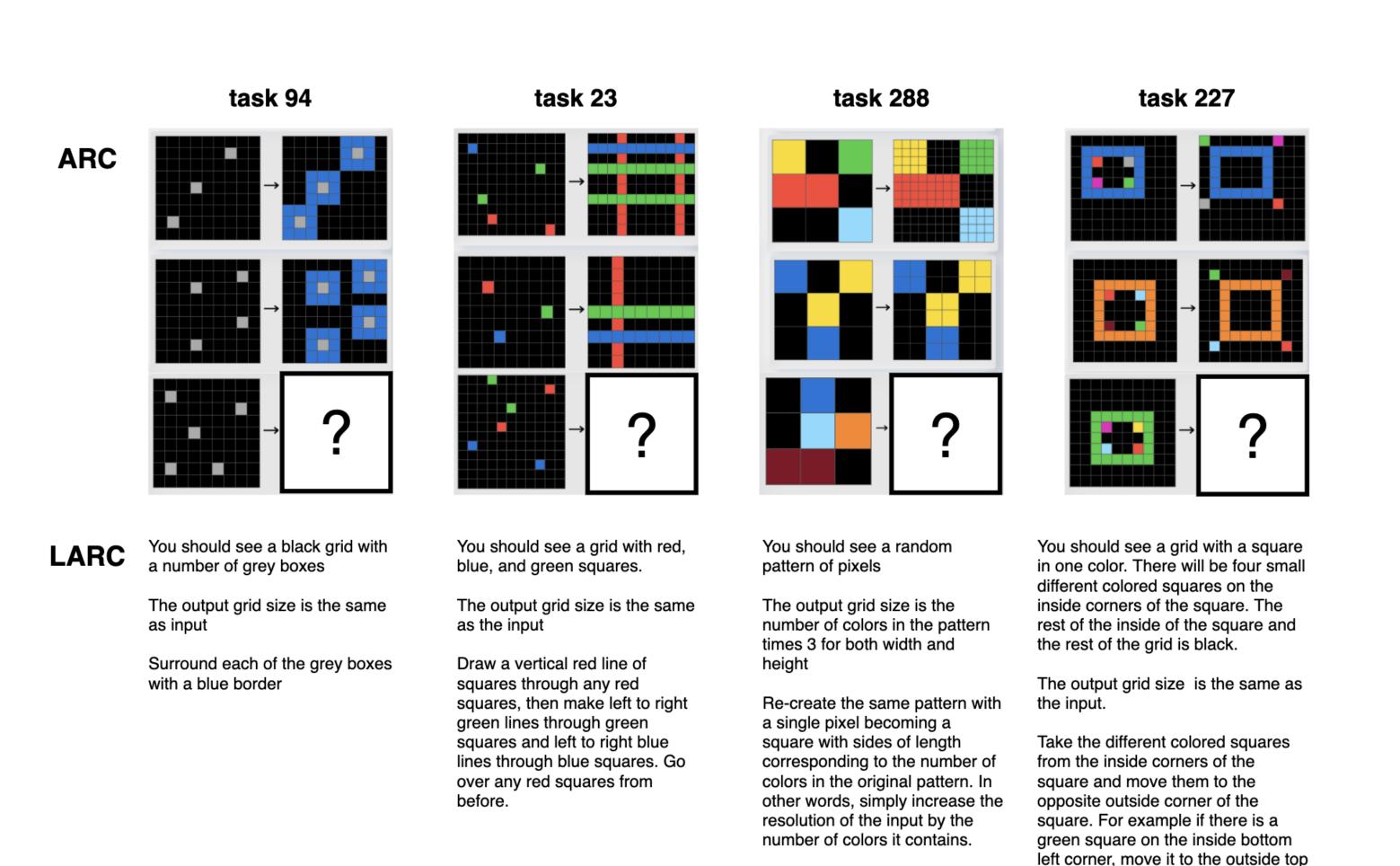
Words used by humans to describe successfully built ARC tasks, sorted by their frequency in the corpus.

[1] Chollet, François. "On the measure of intelligence." arXiv preprint arXiv:1911.01547 (2019)

[2] Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson.

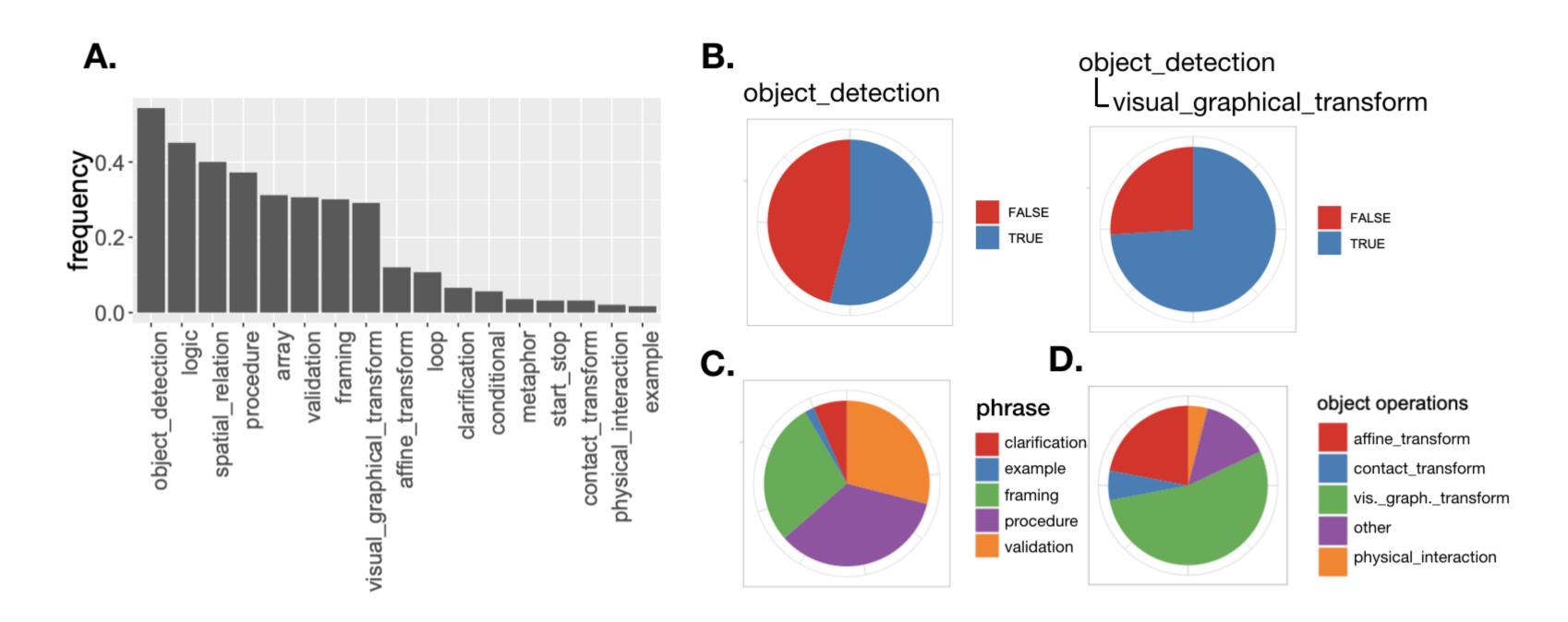
Origins of knowledge. Psychological review, 99(4):605, 1992

The LARC Data-set



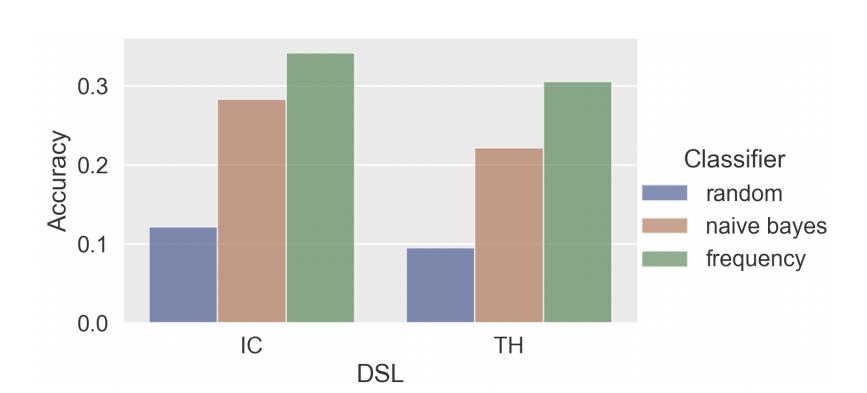
Tasks from the ARC dataset (above) and the corresponding natural programs from LARC (below). For a set of given input-output examples (2 shown here for each task), the ARC task is to generate the correct output on a new input. In LARC, a **describer** is tasked to describe the rule present in the input-output examples so that a different person, a **builder**, can generate the correct output using the description alone.

Topic Analysis



More than half of the phrases described objects, of which, 75% described spatial relations. The fraction of framing, procedure, and clarification occur at roughly the same frequency. More than a half of phrases referred to object modifications (such as extending, colouring, merging or dividing an object), and about a quarter described an affine transform (such as translating, rotating, or scaling).

Synthethiser performance



The predictive accuracy of each classifier predicting the primitives using the natural language descriptions.

Language-guided program search

Model	Solved(720s)	Solved(1hr)	Solved(8hr
BaseDSL	11/400	26/400	57/400
Bigram	33/400	44/400	58/400
CNN	36/400	43/400	64/400
LanT5-N	29/400	41/400	64/400
LanT5P	59/400	66/400	70/400

Models:

right corner. The rest of the squares

will be black.

- BaseDSL: No language initial enumeration from the base DSL
- Bigram: No language fitted bigram baseline
- CNN: No language CNN task encoder for neural search
- LanT5-N: Language T5 language encoder for neural search
- LanT5P: Language T5 + pseudo-annotation training

Tasks solved by language and non-language-guided models, with learned search and learned libraries. Iterative distant supervision training (n=5 iterations). We initialize models with 720s of initial search, yielding 11 initial solved tasks; 1 hour and 26 solved tasks; and 8 hours and 57 solved tasks.

[3] Johnson, A., Vong, W. K., Lake, B. M., Gureckis, T. M. (2021). Fast and flexible: Human program induction in abstract reasoning tasks. arXiv preprint arXiv:2103.05823.

[4] GitHub: https://github.com/top-quarks/ARC-solution