CodeIt: Self-Improving Language Models with Prioritized Hindsight Replay

Natasha Butt [1 2] **Blazej Manczak** [2] **Auke Wiggers** [2]
Corrado Rainone [2] **David W. Zhang** [2] **Michaël Defferrard** [2] **Taco Cohen** [2 3]

Abstract

Large language models are increasingly solving tasks that are commonly believed to require human-level reasoning ability. However, these models still perform very poorly on benchmarks of general intelligence such as the Abstraction and Reasoning Corpus (ARC). In this paper, we approach ARC as a programming-by-examples problem, and introduce a novel and scalable method for language model self-improvement called Code Iteration (CodeIt). Our method iterates between 1) program sampling and hindsight relabeling, and 2) learning from prioritized experience replay. By relabeling the goal of an episode (i.e., the target program output given input) to the realized output produced by the sampled program, our method effectively deals with the extreme sparsity of rewards in program synthesis. Applying CodeIt to the ARC dataset, we demonstrate that prioritized hindsight replay, along with pre-training and data-augmentation, leads to successful inter-task generalization. CodeIt is the first neuro-symbolic approach that scales to the full ARC evaluation dataset. Our method solves 15% of ARC evaluation tasks, achieving state-of-the-art performance and outperforming existing neural and symbolic baselines. Our code is available at https://github.com/Qualcomm-AI-research/codeit.

1. Introduction

The Abstraction and Reasoning Corpus (ARC) is a general artificial intelligence benchmark targeted at both humans and AI systems (Chollet, 2019). ARC is a challenging benchmark because it contains few-shot example tasks that assume access to the four innate core knowledge systems: objects, actions, number, and space (Spelke & Kinzler, 2007). It was designed to require no knowledge outside of these priors, and so the massive memorization capability of pre-trained language models is of limited use for this problem. Humans are able to solve 80% of (a random subset of) ARC tasks in user studies (Johnson et al., 2021), whereas state-of-the-art neural approaches based on GPT-4 solve only 12% of evaluation tasks (Gendron et al., 2023).

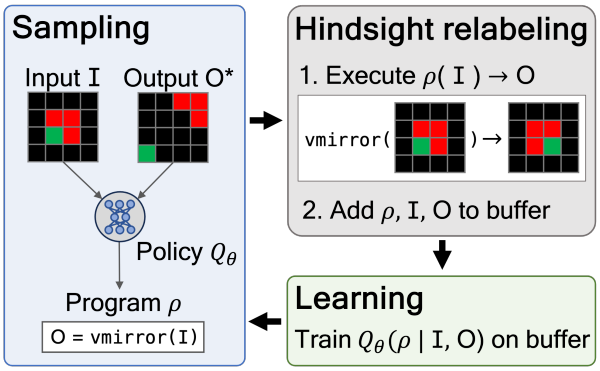


Figure 1. An overview of Code Iteration. In the sampling stage, programs ρ are sampled from the policy Q_θ conditioned on input-output pairs. The program may not produce target output O^* given I, so we use hindsight relabeling: we execute the program, and add the program ρ, inputs I, and *realized* outputs O to the buffer. In the learning stage, we train the policy on samples from the buffer.

Each ARC task consists of a number of *demonstration examples*, each consisting of an input and output grid, and one or more test inputs for which the corresponding output must be predicted (see Figure 2). Effective agents use abstractions related to the four core knowledge systems, generalize from demonstration to test examples, and generalize between tasks. For example, an agent may infer that adjacent cells (space) of the same color value (number) form an object. An agent may also infer that multiple objects sometimes attract or repel (action). Using these abstractions to reason about the value of the test output, an agent may generalize from the demonstration examples to the test example.

Existing approaches to ARC can be classified as either neu-

[1]University of Amsterdam [2]Qualcomm AI Research. Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc. [3]Work was completed while an employee at Qualcomm Technologies Netherlands B.V.. Correspondence to: Natasha Butt <n.e.butt@uva.nl>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

ral (Gendron et al., 2023; Mirchandani et al., 2023), meaning they directly predict output grids using a neural network, or (neuro-) symbolic (Ainooson et al., 2023; Ferré, 2021; 2023), meaning they first predict a program or other symbolic representation of the mapping between input and output grids, before using it to generate the output grids. Through the use of a well-designed *domain-specific language* (DSL), the symbolic methods can be endowed with prior knowledge analogous to the core knowledge systems found in humans. By combining neural networks and symbolic representations like programs, the system can leverage both prior knowledge and data to solve the ARC tasks.

However, the most effective existing methods, whether neural or symbolic, fail to use experience to generalize between tasks. We propose using Expert Iteration (ExIt) (Anthony et al., 2017) to incorporate experience. ExIt methods do this by alternating between two phases: gathering data with an (often expensive) exploration policy, and improving the policy by training on the newfound experiences. Instead of performing ExIt in the grid space, we take a neuro-symbolic approach and train our model to learn to write programs. This brings us closer to the system that emulates general fluid intelligence described by Chollet (2019): by incorporating new experiences in the form of abstractions.

Recent ExIt approaches employ self-improving language models (Gulcehre et al., 2023; Aksitov et al., 2023; Wang et al., 2023c) to replace the expensive expert by sampling from a language model policy and reward-based filtering, saving only trajectories that obtain high reward. This allows them to scale well and benefit from knowledge already captured in the policy. These methods prove effective on program synthesis tasks with natural language specifications (Singh et al., 2023) and code specifications (Haluptzok et al., 2022). However, when solving ARC, agents start ExIt with poor prior knowledge about the search space, as the task is out-of-distribution. Finding a correct program is challenging: positive rewards are extremely sparse. As a result, these methods are sample inefficient in the context of ARC, and programming-by-examples more generally. To enable learning in sparse-reward settings, hindsight relabeling (Andrychowicz et al., 2017) creates artificial expert trajectories post-hoc, and methods that combine ExIt and this technique have improved sample efficiency (Gauthier, 2022; Butt et al., 2022). However, since the relabelled data distribution is constantly changing, there is risk of catastrophic forgetting (French, 1999).

In this work, we introduce a novel, scalable expert iteration method for sparse reward settings that does not suffer from catastrophic forgetting. Our method, which we call Code Iteration or *CodeIt* for short, iterates between 1) a sampling and hindsight relabeling stage and 2) a learning stage with prioritized experience replay. We show a visualization in Figure 1. This iterative procedure thus allows us to automatically generate new data without human intervention. Unlike current self-improvement approaches that perform sampling and filtering (Singh et al., 2023), CodeIt learns from all program samples, improving sample efficiency. By prioritizing training on experiences that solve real tasks, we ameliorate the risk of catastrophic forgetting.

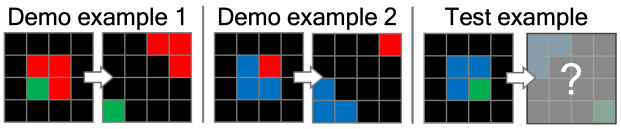


Figure 2. A simplified ARC task. Given two demonstration input-output pairs, the goal is to determine the output grid for the test example, in three attempts or fewer. The size of the grids and the number of demonstration and test examples differs across tasks.

CodeIt solves 59/400 ARC evaluation tasks, achieving state-of-the-art performance by learning from experiences in the form of abstractions and generalizing to new tasks. We analyze the programs discovered by CodeIt and find that these are on average shorter and use different primitives compared to our custom symbolic baselines. Furthermore, after finding an initial solution, CodeIt continues to improve it over time; shorter solutions are found in 53% of solved ARC tasks, highlighting the ability to perform program refinement. We perform careful ablations to better understand the impact on task performance of key components: ExIt, prioritized hindsight replay, and prior knowledge.

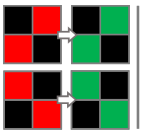
2. Method

We approach ARC as a programming-by-examples problem: for a given set of tasks that we call the *search set*, we aim to find programs that correctly match inputs with their respective outputs, and we do so by training a *policy* to produce programs when shown demonstration examples. This is achieved by iterating between two stages: 1) writing programs using a policy and applying hindsight relabeling, and 2) learning from the programs and their input-output examples. We first describe key design choices below, and then explain the iterative procedure.

2.1. Design choices

Programming language We restrict our programming language to the open source domain specific language (DSL) of Hodel (2023). Although various open source DSLs for ARC exist, Hodel designed their DSL using only the ARC training split, whereas some authors incorporate priors from the ARC evaluation split into their DSLs (Icecuber, 2020).

Hodel's DSL contains grid manipulation functions (e.g., `vmirror` or `hmirror`, which mirror the grid along the



2x2 bg=0 1=0,1 1,0|2x2 bg=0 1=1,0 0,1
2x2 bg=0 2=0,1 1,0|2x2 bg=0 2=1,0 0,1

Figure 3. Sparse grid representation of a simplified ARC task.

vertical or horizontal axis), `fill` functions that replace all pixels of a certain color, and functions that return locations of specific pixel groups. See Appendix B.4 for details on the DSL and more example primitives, and see Hodel (2023) for discussion on the DSL's primitives and capability.

Policy Our choice of policy is a pretrained encoder-decoder Large Language Model (LLM). We use the 220 million parameter CodeT5+ (Wang et al., 2023b) model and its default tokenizer, which are pretrained on a diverse set of programming tasks. We input the demonstration examples to the encoder, and let the decoder generate the corresponding program. If necessary, demonstration examples are truncated to fit in the encoder context window.

Grid representation In order to condition the language model policy on input-output grids, we represent them as text. Instead of encoding the grid as a 2-dimensional array, we use an object-centric text representation. Each color is encoded as an integer, and for each color in the grid we list all the grid cells with that color as $[x, y]$ coordinates. Since the majority of cells belong to the background color, this procedure significantly reduces the number of tokens required to encode the grid (see Figure 16 in Appendix A.3). An example of the sparse grid representation is shown in Figure 3. This object-centric text representation, similar to the one of Xu et al. (2023), works well for sparse grids and is human-interpretable.

2.2. The Code Iteration Algorithm

We initialize the policy network by training on ground truth data. We then start CodeIt, iterating between *sampling and hindsight relabeling* and *learning*. We refer to one full pass of sampling and learning as a *meta-iteration*. We show the procedure in Fig. 1, and explain each stage in more detail below. For pseudocode, see Appendix A.1.

Initialization We start from a dataset of ARC training tasks and solution programs written in the domain-specific language (DSL) of Hodel (2023), which we call the *training set*. This dataset is expanded by randomly mutating programs (for details of this procedure, see Appendix A.2), resulting in an *augmented training set*.

The initial dataset augmentation step serves multiple purposes. Mixing in mutated programs acts as a form of data augmentation, and is a common approach in policy improvement for program synthesis (Ellis et al., 2020; Fawzi et al., 2022). Before experiences are sampled from the policy, the model can already learn the DSL syntax, which can be challenging if the training set is small. It also enables the model to learn how to interpret the task demonstration examples before we begin iterative learning, improving the quality of our policy samples in early meta-iterations.

Sampling and hindsight relabeling In the sampling stage, we obtain new programs using the policy Q_θ. Let the *search set* be the set of tasks for which we want to find a corresponding program. For each task in the search set, we convert the demonstration examples' input I and target output O^* from grid to text representation, encode these using the policy, and then autoregressively decode a program: $\rho \sim Q_\theta(\rho|I, O^*)$. We then run the obtained program on the input grids. If the program is syntactically incorrect or the runtime is too high, we discard it. Otherwise, we obtain program outputs $O = \rho(I)$, and can add a new triplet to a replay buffer: the program ρ, the demonstration inputs I, and the realized outputs O (which may or may not match the target outputs O^*). In each sampling stage we repeat this procedure n_ρ times per task, where n_ρ is a hyperparameter.

Replacing the target output by the realized one is a form of hindsight experience replay (Andrychowicz et al., 2017), and ensures that we obtain an experience every time we find a syntactically correct program, thereby preventing stagnation of the buffer. Although these programs may not solve the tasks we are interested in, they are always valid in terms of syntax and semantics (correctly mapping $\rho(I) \to O$). They can therefore be used to teach the policy about program syntax and program behaviour, which may lead to positive transfer to the search set. We emphasize that we never add test examples nor performance on the test examples to our buffer, as one should not have access to their target output grid during sampling.

Learning During the learning stage, the policy Q_θ is trained on experiences sampled from the buffer, the training set and the augmented training set. These experiences consist of input grids I, output grids O and the corresponding program ρ. The training objective is then a straightforward negative log-likelihood objective:

$$\mathcal{L}(\rho, I, O) = -\log Q_\theta(\rho|I, O). \tag{1}$$

We keep only a single copy of the policy network, and continue to update it during each learning stage. Since we do not compare the policy with its past versions there is no guarantee for improvement. Although continual updates could lead to worse performance in the next iteration, we find this is not a problem in practice.

By default, we perform prioritized sampling from the replay

buffer (Schaul et al., 2015). For each experience, the priority is proportional to the percentage of demonstration outputs equal to program outputs. This means that programs that solve real ARC tasks' demonstration examples are sampled more often than programs for hindsight-relabeled tasks.

3. Experiments

In this section, we aim to demonstrate the efficacy of CodeIt, and break down how much different components of the method contribute to the performance. We first tuned hyperparameters on a custom training and validation split (for a description of these parameters and details, see Appendix B). Using these hyperparameters, we benchmark our method on the ARC evaluation split and compare against previous state-of-the-art methods. Finally, we ablate the importance of individual components of CodeIt.

We define *demonstration performance* as the percentage of solved demonstration examples on a given task. We first sort solution programs by demonstration performance, and then by program length, favoring shorter programs. We evaluate the top three programs on the set of test examples. Following ARC evaluation procedure, if at least one of these three programs maps all test example inputs to outputs, the task is solved and *test performance* is 1. We emphasize that the ExIt procedure only makes use of demonstration examples, and that we use test performance for final evaluation only.

Custom baselines We use a random baseline that samples programs line-by-line. At the start of each line, we sample a primitive function from the DSL, then sample arguments given its expected input types. When a variable of type "grid" is created, we end the program with probability 0.8, otherwise we add another line to the program.

We also use a mutation-based baseline. This is a more advanced procedure, designed with the DSL in mind. At every meta-iteration, it mutates the set of training programs provided by Hodel (2023). We use two variations: "d_1" mutates only the initial training set, and "d_∞" can augment newfound programs as well. We provide the exact algorithm in Appendix A.2.

For all three baselines, we sample $n_m = n_\rho \cdot n_{tasks}$ programs per meta-iteration. Here, n_ρ is the desired number of programs per meta-iteration per task, and n_{tasks} the total number of tasks in the population. To strengthen these baselines, we exhaustively evaluate each found program on all inputs in the search set, and check the outputs against ARC output grids.

Baselines from literature We include approaches from literature as baselines as well. A direct comparison is sometimes difficult, as not all baselines apply their method to the full ARC evaluation set: for example, Kolev et al. (2020) and Alford et al. (2021) focus only on a subset of ARC. Additionally, some symbolic methods design a DSL based on both ARC training and evaluation sets and report results on a hidden test set (Icecuber, 2020). We therefore only compare to approaches that report scores on the full ARC evaluation set.

Ainooson et al. (2023) and Ferré (2023) both run a search procedure for a custom DSL on the full set. As Ainooson et al. (2023) report the highest performance the full ARC evaluation set, this is our main symbolic baseline. Although Mirchandani et al. (2023) and Gendron et al. (2023) use a different evaluation protocol, we include these as our main neural baseline, as they are based on powerful LLMs (text-davinci and GPT-4).

3.1. Setup

We initialize our training set with the 400 examples from the ARC training split and the associated solution programs provided by Hodel (2023). We also sample 19,200 programs as additional training data via the mutation procedure outlined in Appendix A.2. We use the programs that are syntactically correct to initialize the augmented training set. We use the 400 ARC evaluation examples as our search set.

In the sampling stage of each meta-iteration, we use temperature sampling with temperature $\tau = 0.95$, and sample up to $n_\rho = 24$ programs per task. This encourages exploration and, as a result, increases the diversity of data added to the replay buffer. We reject policy-sampled programs if they are syntactically incorrect, or if they run for more than 0.25 seconds per program line. All valid programs are added to the replay buffer.

In each learning stage, we start by sampling a set of experiences from the buffer under the distribution given by the priorities. Each meta-iteration, we sample $r_t = 10,000$ experiences from the concatenation of the train set and the augmented train set, and $r_p = 90,000$ experiences from the buffer. The resulting set is used for 1 epoch of training. For a full list of hyperparameters, see Table 3 in the Appendix.

3.2. Main results on ARC eval set

In Figure 4, we show performance as a function of the number of sampled programs, for CodeIt, our custom baselines, Ainooson et al. (2023) and Ferré (2023). We show *cumulative performance* here, which means that any program in the buffer or augmented train set is considered a solution candidate. For the mutation baselines, we see a rapid performance increase followed by stagnation. In comparison, CodeIt takes several meta-iterations to start generating solutions outside of the augmented train set and then performance rapidly increases. CodeIt quickly outperforms

Method	ARC Train Set	ARC Eval Set	ARC Eval 412
Ferré (2021)	29 / 400	6 / 400	-
Ainooson et al. (2023) MLE	70 / 400	17 / 400	-
Ainooson et al. (2023) brute force	104 / 400	26 / 400	-
Ferré (2023)	96 / 400	23 / 400	-
Mirchandani et al. (2023) text-davinci-003	56 / 400*	27 / 400*	-
Gendron et al. (2023) GPT-4	-	-	49 / 412*
Mutation d_1 baseline	-	42 / 400	39 / 412*
Mutation d_∞ baseline	-	38 / 400	36 / 412*
Random baseline	-	6 / 400	7 / 412*
CodeIt	-	**59 / 400**	**59 / 412***

Table 1. Main results on ARC eval set. The evaluation metric is pass@3 by default, * indicates pass@1. To enable comparison to related work of Gendron et al. (2023), we also include pass@1 performance on the ARC Eval set with 412 examples. Our method outperforms all previous baselines. More details on the ARC splits and evaluation procedures can be found in Appendix A.4.

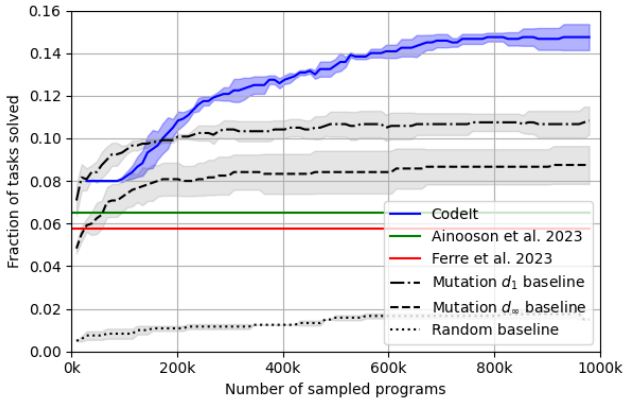


Figure 4. Cumulative performance as function of number of sampled programs for CodeIt and various baselines, showing mean and standard deviation of three runs for CodeIt and custom baselines.

the mutation baseline, indicating that it indeed finds higher-quality samples to train on.

We report final performance of CodeIt after 100 meta-iterations, and the performance of various baselines, in Table 1. To enable comparison to Gendron et al. (2023), we include results on the "ARC Eval 412" set, which treats each test example in the ARC evaluation set as a separate task. Our approach outperforms symbolic approaches (Ainooson et al., 2023; Ferré, 2021; 2023), but also neural approaches based on large language models (Gendron et al., 2023; Mirchandani et al., 2023), achieving state-of-the-art performance on the ARC evaluation set.

For context, we show a solution written by CodeIt for an example task in Figure 5. To further illustrate the differences between the programs found by CodeIt and the mutation baselines, we analyze solutions found by each method in Appendix E.3, including a qualitative comparison in Table 4. One finding is that there are 29 tasks for which CodeIt and the mutation baseline both find a solution, but that there are 23 tasks for which only CodeIt finds a solution, versus 13 for the mutation baseline. For the tasks that both methods solve, CodeIt finds shorter programs on average and uses different primitives.

In Appendix E.4, we observe CodeIt refines its initial solution for 53% of solved tasks, producing a shorter solution in a later meta-iteration. Moreover, in Appendix E.1, we analyze failure cases on the customer validation set and observe that CodeIt does not solve tasks with solution programs longer than 11 lines but often learns to use their primitives. Further, in Appendix E.2, we look at the DSL primitives that CodeIt learns and find that some primitive types are learned quicker than others. Finally, in Appendix C, we find that CodeIt appears to perform best on tasks related to object interactions and worst on numerical or logic based tasks by analyzing performance on the ConceptARC dataset (Moskvichev et al., 2023).

3.3. Ablations

In Figure 6 and 7, we report cumulative performance and policy performance over time for CodeIt and all ablations. In all cases, we initialize the method with the ARC train set, and use the ARC evaluation set as search set. We show the results of ablations at the end of training in Table 2. We also perform a scaling study with different model sizes, results are shown in Appendix D.

A1: No ExIt This ablation removes policy feedback, to isolate the contribution of Expert Iteration. In every meta-iteration, instead of populating the buffer with policy samples, we take the programs generated in that meta-iteration of the mutation d_1 baseline. For each program, we randomly select a task from the search set and perform hindsight re-

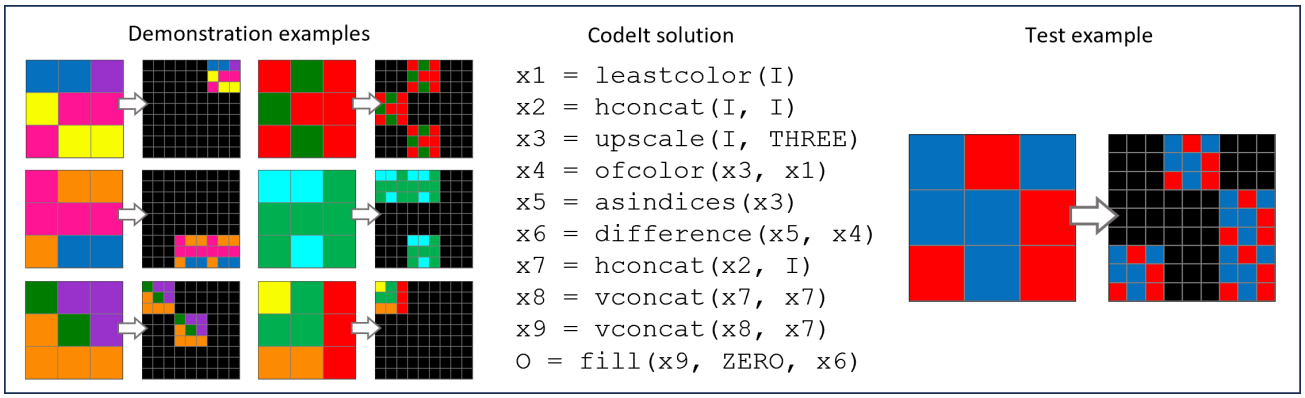


Figure 5. ARC evaluation task 48f8583b and the solution program found by CodeIt.

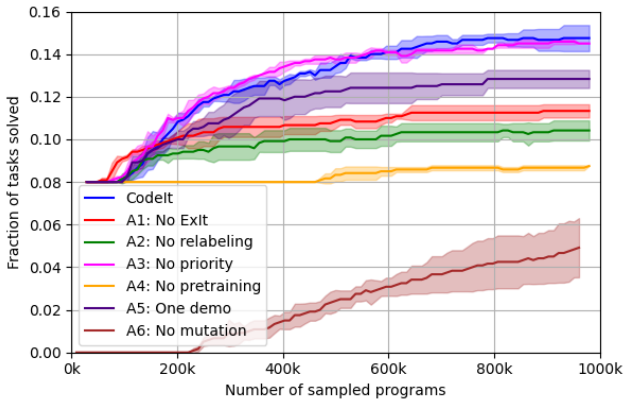


Figure 6. Cumulative performance as function of number of sampled programs for CodeIt and ablations, for three random seeds. For cumulative performance, all programs in the augmented train set and buffer are candidate solutions.

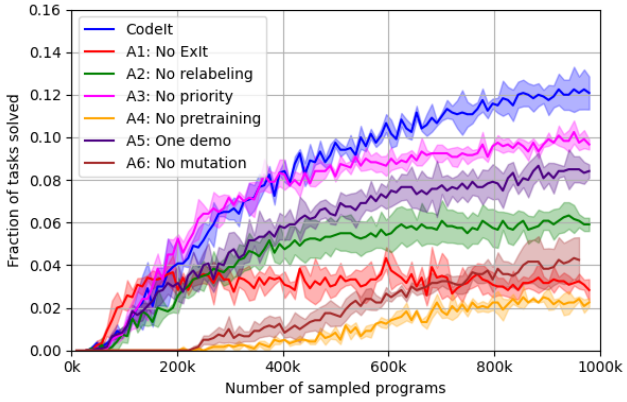


Figure 7. Policy performance per meta iteration as function of number of sampled programs for CodeIt and ablations, for three random seeds. For policy performance, only programs output by the policy in the *current* meta-iteration are candidate solutions.

labelling, adding the program, input, output triplet to the buffer. We sample $r_p + r_t = 100,000$ experiences from the concatenation of the train set, the augmented train set and the buffer at each meta-iteration for learning. We see that A1 outperforms the mutation baseline, which means supervised learning from mutation experiences alone does lead to some inter-task generalization. However, cumulative performance is substantially lower than CodeIt. This highlights the importance of policy feedback.

A2: No relabeling We test the effect of hindsight relabeling by only adding experiences to the buffer if the program produces the correct output for all demonstration examples. We train on all experiences in the buffer without prioritized sampling. Although performance increases in early meta-iterations, A2 stagnates after around 30 meta-iterations, indicating that data generated by sampling and filtering alone is not sufficient. Sampling and hindsight relabeling (CodeIt) performs better than sampling and filtering (A2).

A3: No priority To test the hypothesis that prioritized sampling ameliorates catastrophic forgetting, we draw experiences uniformly from the buffer in the learning stage. A3 leads to a small reduction in cumulative performance, but a large reduction in policy performance, indicating that the policy indeed forgets important experiences. Prioritized sampling results in better retention of knowledge.

A4: No pretraining To identify whether our pre-trained policy contains beneficial prior knowledge, we randomly reinitialize the policy's weights at the start of CodeIt. Policy performance shows that performance improvement is much slower. Moreover, inter-task generalization begins later, as shown by the cumulative performance, which only starts increasing after around 50 meta-iterations. Despite the expected slowdown, it is encouraging to see that CodeIt does seem to be able to bootstrap from random weights.

A5: One demo We investigate use of the task representation by decreasing the number of demonstration examples shown to the policy. This results in a strong decrease in both cumulative and policy performance. This indicates CodeIt forms abstractions over multiple demonstration examples.

A6: No mutation In this ablation, we omit the mutation-based training data augmentation step. We observe that taking out mutation-based bootstrapping results in slower training, although performance does increase over time and does not stagnate. We therefore conjecture that mutation-based augmentation is not strictly required but still useful.

4. Related work

4.1. Abstraction and Reasoning Corpus (ARC)

Various works have applied program synthesis approaches to subsets of the ARC dataset. Xu et al. (2022) proposes to represent grids as graphs, and applies logical programs to the graph nodes, solving 63 of 160 tasks. Kolev et al. (2020) apply a Differentiable Neural Computer to ARC, solving 78% of tasks with grids of size 10×10 and smaller. Alford et al. (2022) applies DreamCoder (Ellis et al., 2020) and execution-guided program synthesis, solving 22 of 36 considered tasks. Park et al. (2023) first collects human feedback, then performs behavioral cloning for a subset of ARC tasks using a decision transformer (Chen et al., 2021). However, none of these methods are applied on the full ARC evaluation set, typically due to poor scaling behavior.

The few works that do scale to the full evaluation set tend to solve each task in isolation. Ferré (2021) and followup work Ferré (2023) design a custom DSL and perform a fast search for each task. Ainooson et al. (2023) also designs a custom DSL and obtains best performance with a brute-force search, solving 36 of 400 evaluation tasks. Mirchandani et al. (2023) and Gendron et al. (2023) demonstrate that a pretrained language model with custom tokenizer will output the correct grid after being shown multiple input-output pairs, solving 27 of 400 and 49 of 412 evaluation tasks respectively. Wang et al. (2023a) augment this approach by generating hypotheses in multiple rounds, although they only show performance on a subset of the ARC training set due to the high monetary cost of querying the language model. In this work, we design a scalable ExIt approach that combines a smaller language model with the higher-level abstraction of a DSL. We also ensure that our approach incorporates experience to benefit from generalization between tasks.

Various unpublished approaches exist too, including submissions to ARC challenges as well as a Kaggle competition. These competitions use a private leaderboard, not revealed to participants. This means participants often use the public ARC evaluation set for training or DSL design purposes. For example, the winner of Kaggle 2020 comments that searching in a DSL designed using the training set resulted in low performance, and higher performance was reached after conditioning the DSL on the evaluation tasks (Icecuber, 2020). This makes direct comparisons to methods evaluated on the evaluation set difficult. For reference, we include a summary of competition results in Appendix F Table 7, however, note that this summary reports performance on the hidden test set, and that competition results cannot be directly compared to this work and the literature.

4.2. Expert Iteration

Expert iteration (ExIt) (Anthony et al., 2017) consists of a policy-guided search stage that gathers new experiences, and a learning stage that improves the policy by imitation learning. Commonly used experts tend to be powerful and computationally intensive tree search algorithms such as Monte Carlo Tree Search (Kocsis & Szepesvári, 2006) and greedy search (Daumé et al., 2009). ExIt has achieved super-human performance include games (Silver et al., 2016; 2018; Anthony et al., 2017) and combinatorial problems such as bin-packing (Laterre et al., 2019). Related work that employ hindsight relabeling in expert iteration are Aygün et al. (2021), Butt et al. (2022) and Gauthier & Urban (2022).

Applications of ExIt for programming-by-examples (Mankowitz et al., 2023; Ellis et al., 2020) are most relevant to CodeIt. Mankowitz et al. (2023) consider one task only: writing a fast sorting algorithm. For this problem, inter-task generalization is thus not as important. DreamCoder (Ellis et al., 2020) is most related to our work, since this ExIt method is applied to multiple programming-by-examples tasks. DreamCoder uses a continually growing DSL to store abstractions, and a computationally intensive search procedure. Instead, CodeIt uses the model to store distilled knowledge, and generates experiences via sampling from the model. Furthermore, DreamCoder filters solutions based on correctness whereas CodeIt uses hindsight relabeling and prioritized experience replay.

4.3. Self Improving Large Language Models

Previous work showed that learning from synthetic data is a viable strategy for theorem proving (Wang & Deng, 2020) and programming-by-examples (Balog et al., 2017; Devlin et al., 2017; Bunel et al., 2018; Parisotto et al., 2017; Polosukhin & Skidanov, 2018; Zohar & Wolf, 2018), often training a model from scratch. Instead, finetuning pre-trained large language models (LLMs) on synthetic data enables knowledge transfer due to the prior domain knowledge captured in their weights (Butt et al., 2022). Recently, methods that use LLMs to synthesize training data have shown successes in general domains including theorem proving (Polu et al., 2022), question answering (Zelikman et al.,

Method	initial policy weights	# demo examples	# policy samples	policy only perf.	cumulative perf.
CodeIt	CodeT5	≤ 10	24	49/400	59/400
A1: No ExIt	CodeT5	≤ 10	0	13/400	45/400
A2: No relabeling	CodeT5	≤ 10	24	24/400	42/400
A3: No priority	CodeT5	≤ 10	24	38/400	58/400
A4: No pretraining	Random	≤ 10	24	9/400	35/400
A5: One demo	CodeT5	≤ 1	24	34/400	51/400
A6: No mutation	CodeT5	≤ 10	24	17/400	20/400

Table 2. ARC evaluation performance of CodeIt ablations.

2022; Aksitov et al., 2023), mathematical reasoning (Ni et al., 2023), machine translation (Gulcehre et al., 2023), language-to-code generation (Zhou et al., 2023; Singh et al., 2023) and code-to-code generation (Haluptzok et al., 2022). We demonstrate in this work that such an approach can be applied to the challenging ARC domain as well.

5. Discussion

Various factors make ARC uniquely challenging for learning-based approaches, for example the limited amount of training data, and the complexity of individual tasks. Another issue is that programs may differ in number of demonstration examples and input dimensionality, which requires agents to reason about concepts at different scales. In this work, we show that an expert iteration based approach can learn to solve 59/400 unseen ARC tasks. Here, we provide intuition for why CodeIt works well on this benchmark.

Ablations showed that hindsight relabeling has a large effect on performance. Many expert iteration approaches rely on the emergence of a curriculum of increasingly difficult tasks, even creating a curriculum by comparing the current agent to past versions of itself (Silver et al., 2016; Fawzi et al., 2022) or reward shaping (Laterre et al., 2019; Gulcehre et al., 2023). Hindsight relabeling forms an implicit curriculum (Andrychowicz et al., 2017): initially we collect easy tasks that can be solved in few lines of code, while later on, programs become more complex. This is useful for ARC, where obtaining even one solved task is challenging. As relabeling adds many programs to the buffer, including some that are further away from the target tasks, we used prioritized sampling to avoid catastrophic forgetting.

A potential limitation of CodeIt is that for ARC, it relies on hand-designed components: a domain specific language (DSL), access to an interpreter for automatic evaluation, and an initial set of ground truth programs. While we do benefit from Hodels expert-designed DSL, we also showed that a neuro-symbolic approach (ablation A1) outperforms a symbolic approach (the mutation baseline), indicating that both DSL and learning contribute to performance. Further, CodeIt outperforms both, indicating that ExIt compounds this effect. We also use a pretrained LLM and mutation procedure to speed up training, but ablations showed that training is possible even without these, albeit at a slower pace. Nevertheless, approaches that can start learning tabula rasa, or form their own DSL (Ellis et al., 2020) remain an important area of research.

For the ARC dataset, it is currently beneficial to incorporate both prior knowledge (via a DSL or pre-trained LLM) and experience (via expert iteration). Chollet (2019) defines the intelligence of a system as "a measure of its skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty". Chollet poses that, if two systems are initialized with the same prior knowledge and go through the same amount of experience with respect to a set of unseen tasks, the more intelligent system will combine prior knowledge and its experience more efficiently, solving more tasks.

Although many existing approaches incorporate prior knowledge through a programming language or DSL (Ainooson et al., 2023; Ferré, 2023), a pre-trained large language model (Gendron et al., 2023; Mirchandani et al., 2023), or both (Wang et al., 2023a), they cannot incorporate new experience, and therefore do not benefit from inter-task generalization. Alford (2021) proposes an expert iteration method that does learn from experience, but it does not scale well nor benefit from prior knowledge in its policy. We pose that CodeIt is the more effective expert iteration method due to its use of scalable components: pre-trained language models, likelihood-based training, and running programs in interpreters. There is also an implicit relationship between *computational* efficiency and experience: since CodeIt's policy learns on the ARC domain, it is possible to use a much smaller language model than for example Gendron et al. (2023), who use GPT-4 as a policy. This is consistent with LLM literature showing that high quality training data with a curriculum enables smaller LMs to compete with much larger ones on coding tasks (Gunasekar et al., 2023).

6. Conclusion

We introduce a novel and scalable method for self-improving language models, *CodeIt*, that uses prioritized hindsight replay. CodeIt achieves state-of-the-art performance on the Abstraction and Reasoning Corpus (ARC) compared to symbolic and neural baselines, solving 59 of 400 evaluation tasks. Ablations show that hindsight relabeling leads to improved sample efficiency resulting in a 40% improvement in performance. We also find that prioritizing important experiences during training ameliorates catastrophic forgetting. Additionally, we observe that CodeIt is able to refine solutions over time, identifying a shorter program for 53% of solved tasks in later iterations. The results demonstrate that our self-improving language model is capable of reasoning in the program space and generalizing between tasks. For the challenging ARC benchmark, both scalability and learning from experience prove to be key components for success.

Acknowledgements

We thank Michael Hodel for the creation of their DSL for solving ARC tasks, available at https://github.com/michaelhodel/arc-dsl.

Impact Statement

This work presents a method for training a language model based policy on programming-by-example problems where data is scarce. Our approach learns to write programs in iterations, and receives feedback from an interpreter.

On one hand, large language models have potential negative societal impacts, exacerbating biases present in training data. This can be especially harmful if datasets are small, as is the case in our work. Additionally, models that generate code can be used to automate writing of malicious code.

On the other hand, enabling the use of large language models on problems where data is scarce opens up application areas where it is currently challenging to deploy them. Models that generate code from input-output specifications are broadly applicable and can be used in many programming settings.

References

Ainooson, J., Sanyal, D., Michelson, J. P., Yang, Y., and Kunda, M. An approach for solving tasks on the abstract reasoning corpus, 2023.

Aksitov, R., Miryoosefi, S., Li, Z., Li, D., Babayan, S., Kopparapu, K., Fisher, Z., Guo, R., Prakash, S., Srinivasan, P., Zaheer, M., Yu, F., and Kumar, S. Rest meets react: Self-improvement for multi-step reasoning llm agent, 2023.

Alford, S. *A Neurosymbolic Approach to Abstraction and Reasoning*. PhD thesis, Massachusetts Institute of Technology, 2021.

Alford, S., Gandhi, A., Rangamani, A., Banburski, A., Wang, T., Dandekar, S., Chin, J., Poggio, T. A., and Chin, P. Neural-guided, bidirectional program search for abstraction and reasoning. *Complex Networks*, 2021.

Alford, S., Gandhi, A., Rangamani, A., Banburski, A., Wang, T., Dandekar, S., Chin, J., Poggio, T., and Chin, P. Neural-guided, bidirectional program search for abstraction and reasoning. In *Complex Networks & Their Applications X: Volume 1, Proceedings of the Tenth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2021 10*, pp. 657–668. Springer, 2022.

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Pieter Abbeel, O., and Zaremba, W. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Anthony, T., Tian, Z., and Barber, D. Thinking fast and slow with deep learning and tree search. May 2017.

ARCathon Leaderboard. https://lab42.global/arcathon/leaderboard/, 2023. Accessed: 2024-30-01.

Aygün, E., Orseau, L., Anand, A., Glorot, X., Firoiu, V., Zhang, L. M., Precup, D., and Mourad, S. Proving theorems using incremental learning and hindsight experience replay, 2021.

Balog, M., Gaunt, A., Brockschmidt, M., Nowozin, S., and Tarlow, D. Deepcoder: Learning to write programs. In *International Conference on Learning Representations*. OpenReview. net, 2017.

Bunel, R., Hausknecht, M., Devlin, J., Singh, R., and Kohli, P. Leveraging grammar and reinforcement learning for neural program synthesis. In *International Conference on Learning Representations*, 2018.

Butt, N., Wiggers, A., Cohen, T., and Welling, M. Program synthesis for integer sequence generation. 2022. URL https://mathai2022.github.io/papers/24.pdf.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling, 2021. URL https://arxiv.org/abs/2106.01345.

Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Daumé, H., Langford, J., and Marcu, D. Search-based structured prediction. *Machine learning*, 75:297–325, 2009.

Devlin, J., Uesato, J., Bhupatiraju, S., Singh, R., Mohamed, A.-r., and Kohli, P. Robustfill: Neural program learning under noisy i/o. In *International conference on machine learning*, pp. 990–998. PMLR, 2017.

Ellis, K., Wong, C., Nye, M. I., Sablé-Meyer, M., Cary, L., Morales, L., Hewitt, L. B., Solar-Lezama, A., and Tenenbaum, J. B. DreamCoder: Growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning. *CoRR*, abs/2006.08381, 2020. URL https://arxiv.org/abs/2006.08381.

Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Mohammadamin, B., Novikov, A., Ruiz, F. J. R., Schrittwieser, J., Swirszcz, G., Silver, D., Hassabis, D., and Kohli, P. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610, 2022. doi: https://doi.org/10.1038/s41586-022-05172-4.

Ferré, S. First steps of an approach to the arc challenge based on descriptive grid models and the minimum description length principle. *arXiv preprint arXiv:2112.00848*, 2021.

Ferré, S. Tackling the abstraction and reasoning corpus (arc) with object-centric models and the mdl principle. *arXiv preprint arXiv:2311.00545*, 2023.

French, R. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3:128–135, 05 1999. doi: 10.1016/S1364-6613(99)01294-2.

Gauthier, T. Program synthesis for the oeis, 2022. URL https://arxiv.org/abs/2202.11908.

Gauthier, T. and Urban, J. Learning program synthesis for integer sequences from scratch, 2022. URL https://arxiv.org/abs/2202.11908.

Gendron, G., Bao, Q., Witbrock, M., and Dobbie, G. Large language models are not strong abstract reasoners, 2023.

Gulcehre, C., Paine, T. L., Srinivasan, S., Konyushkova, K., Weerts, L., Sharma, A., Siddhant, A., Ahern, A., Wang, M., Gu, C., et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., and Li, Y. Textbooks are all you need, 2023.

Haluptzok, P., Bowers, M., and Kalai, A. T. Language models can teach themselves to program better. *arXiv preprint arXiv:2207.14502*, 2022.

Hodel, M. Domain-specific language for the abstraction and reasoning corpus, 2023.

Icecuber. https://www.kaggle.com/competitions/abstraction-and-reasoning-challenge/discussion/154597, 2020. Accessed: 2024-30-01.

Johnson, A., Vong, W. K., Lake, B. M., and Gureckis, T. M. Fast and flexible: Human program induction in abstract reasoning tasks. *CoRR*, abs/2103.05823, 2021. URL https://arxiv.org/abs/2103.05823.

Kaggle Leaderboard. https://www.kaggle.com/competitions/abstraction-and-reasoning-challenge/code, 2020. Accessed: 2024-30-01.

Kocsis, L. and Szepesvári, C. Bandit based monte-carlo planning. In *European conference on machine learning*, pp. 282–293. Springer, 2006.

Kolev, V., Georgiev, B., and Penkov, S. Neural abstract reasoner. *arXiv preprint arXiv:2011.09860*, 2020.

Laterre, A., Fu, Y., Jabri, M. K., Cohen, A.-S., Kas, D., Hajjar, K., Chen, H., Dahl, T. S., Kerkeni, A., and Beguir, K. Ranked reward: enabling self-play reinforcement learning for bin packing. 2019.

Mankowitz, D. J., Michi, A., Zhernov, A., Gelmi, M., Selvi, M., Paduraru, C., Leurent, E., Iqbal, S., Lespiau, J.-B., Ahern, A., et al. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618(7964): 257–263, 2023.

Mirchandani, S., Xia, F., Florence, P., Ichter, B., Driess, D., Arenas, M. G., Rao, K., Sadigh, D., and Zeng, A. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*, 2023.

Mitchell, M., Palmarini, A. B., and Moskvichev, A. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks, 2023.

Moskvichev, A., Odouard, V. V., and Mitchell, M. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain, 2023.

Ni, A., Inala, J. P., Wang, C., Polozov, O., Meek, C., Radev, D., and Gao, J. Learning math reasoning from self-sampled correct and partially-correct solutions, 2023.

Parisotto, E., Mohamed, A.-r., Singh, R., Li, L., Zhou, D., and Kohli, P. Neuro-symbolic program synthesis. In *International Conference on Learning Representations*, 2017.

Park, J., Im, J., Hwang, S., Lim, M., Ualibekova, S., Kim, S., and Kim, S. Unraveling the arc puzzle: Mimicking human solutions with object-centric decision transformer. *arXiv preprint arXiv:2306.08204*, 2023.

Polosukhin, I. and Skidanov, A. Neural program search: Solving programming tasks from description and examples. *CoRR*, abs/1802.04335, 2018. URL http://arxiv.org/abs/1802.04335.

Polu, S., Han, J. M., Zheng, K., Baksys, M., Babuschkin, I., and Sutskever, I. Formal mathematics statement curriculum learning. *CoRR*, abs/2202.01344, 2022. URL https://arxiv.org/abs/2202.01344.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Singh, A., Co-Reyes, J. D., Agarwal, R., Anand, A., Patil, P., Garcia, X., Liu, P. J., Harrison, J., Lee, J., Xu, K., Parisi, A., Kumar, A., Alemi, A., Rizkowsky, A., Nova, A., Adlam, B., Bohnet, B., Elsayed, G., Sedghi, H., Mordatch, I., Simpson, I., Gur, I., Snoek, J., Pennington, J., Hron, J., Kenealy, K., Swersky, K., Mahajan, K., Culp, L., Xiao, L., Bileschi, M. L., Constant, N., Novak, R., Liu, R., Warkentin, T., Qian, Y., Bansal, Y., Dyer, E., Neyshabur, B., Sohl-Dickstein, J., and Fiedel, N. Beyond human data: Scaling self-training for problem-solving with language models, 2023.

Spelke, E. S. and Kinzler, K. D. Core knowledge. *Developmental Science*, 10(1):89–96, 2007. doi: https://doi.org/10.1111/j.1467-7687.2007.00569.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-7687.2007.00569.x.

Wang, M. and Deng, J. Learning to prove theorems by learning to generate theorems. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18146–18157. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d2a27e83d429f0dcae6b937cf440aeb1-Paper.pdf.

Wang, R., Zelikman, E., Poesia, G., Pu, Y., Haber, N., and Goodman, N. D. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*, 2023a.

Wang, Y., Le, H., Gotmare, A. D., Bui, N. D., Li, J., and Hoi, S. C. H. Codet5+: Open code large language models for code understanding and generation. *arXiv preprint*, 2023b.

Wang, Z., Hou, L., Lu, T., Wu, Y., Li, Y., Yu, H., and Ji, H. Enable language models to implicitly learn self-improvement from data, 2023c.

Xu, Y., Khalil, E. B., and Sanner, S. Graphs, constraints, and search for the abstraction and reasoning corpus, 2022.

Xu, Y., Li, W., Vaezipoor, P., Sanner, S., and Khalil, E. B. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *arXiv preprint arXiv:2305.18354*, 2023.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Zhou, A., Yan, K., Shlapentokh-Rothman, M., Wang, H., and Wang, Y.-X. Language agent tree search unifies reasoning acting and planning in language models, 2023.

Zohar, A. and Wolf, L. Automatic program synthesis of long programs with a learned garbage collector. *Advances in neural information processing systems*, 31, 2018.

A. Method and evaluation details

A.1. CodeIt Algorithm

The pseudo code for the CodeIt procedure is portrayed in Algorithm 1.

Algorithm 1 CodeIt Algorithm

Require: Training set D_{train}, search set D_{test}, policy Q
Ensure: Finetuned policy Q, updated replay buffer R, optimal programs set ρ^*
  $D_{\text{augmented_train}} \leftarrow$ EvolveTrainingTasks(D_{train}) {Evolve training tasks}
  Initialize ρ^* as an empty set {Init set of programs that solve tasks in D_{test}}
  for meta_iter $= 1 \rightarrow 100$ **do**
    # Sampling and hindsight relabeling stage
    for task in D_{test} **do**
      $\{\rho\} \leftarrow Q(\rho|\{I, O\})$ {Sample programs for test tasks}
      for each ρ in $\{\rho\}$ **do**
        if SyntactticallyValid(ρ) **then**
          Add $\{\rho, \{(I^{(i)}, \rho(I^{(i)})), \dots\}\}$ to R {Update the replay buffer with hindsight relabeled tasks}
        end if
        for $(I^{(i)}, O^{(i)})$ in task **do**
          if $\rho(I^{(i)}) = O^{(i)}$ **then**
            Add {(ρ, task)} to ρ^* {Update set of programs that solve tasks in D_{test}}
          end if
        end for
      end for
    end for
    # Learning stage
    $D_{\text{sample}} \leftarrow$ SampleFrom($R + D_{\text{augmented_train}} + D_{\text{train}}$) {Sample tasks from the replay buffer}
    Train Q on D_{sample} for 1 epoch {Continual training of the policy}
  end for

Initializing CodeIt Before we start the CodeIt procedure, we expand the training dataset using the first 19,200 mutated tasks from the mutation procedure (see Appendix A.2) used for the mutation d_1 baseline.

A.2. Program and Task Mutation

Mutation procedure To grow a population of mutated programs with task demonstration inputs corresponding to the original training dataset, we follow the procedure outlined in Algorithm 3. This involves mutating a single task, which is described in Algorithm 2. The mutation is carried out with the hyperparameters $\phi_{\text{var}} = 0.25, \phi_{\text{arg}} = 0.5, \phi_{\text{func}} = 0.25$. With respect to naming notation, d_1 reflects a depth of 1, meaning we only mutate programs from the original training set, and d_∞ reflects a depth of infinity, meaning we can mutate previously mutated programs.

The intuitive explanation of the mutation procedure for a single program is as follows. We pick a random line from a program (L2-3). We then replace either a function call with a function with similar output type (L4-7), or we replace an input argument in the function call (L8-11), or we replace the function call but leave its input variables the same (L12-14).

Mutation baseline For our mutation baseline, we sample mutated programs using the mutation procedure outlined above. For all the mutated programs in the evolved task population, we evaluate each program on the tasks in our search set.

A.3. Task Representation

Grid representation We use a compressed grid representation, mainly to reduce the number of tokens needed to represent each grid. We do not use a custom tokenizer. A visualization of the number of tokens is shown in Fig. 16, showing that in almost all cases, the sparse grid representation we use leads to a reduction in the number of needed tokens, especially for larger grid sizes.

Algorithm 2 MutateProgram

Require: Replacement probabilities ϕ_{var}, ϕ_{arg}, ϕ_{func}, program ρ
Ensure: ρ'
  Initialize $\rho' \leftarrow \rho$ {Copy original program}
  $l \leftarrow$ RandomLineFrom(ρ') {Randomly select a line}
  $p \sim U(0, 1)$
  if $p < \phi_{\text{var}}$ **then**
    $f' \leftarrow$ SampleFunctionWithOutputType(GetTypeOfVariable(l))
    $args' \leftarrow$ SampleArgumentsForFunction(f')
    Replace variable definition $f(args)$ in l with $f'(args')$
  else if $p < (\phi_{\text{var}} + \phi_{\text{arg}})$ **then**
    $a \leftarrow$ RandomArgumentFrom(l)
    $a' \leftarrow$ SampleTermOfType(GetTypeOfArgument(a))
    Replace argument a with a'
  else
    $f' \leftarrow$ SampleFunctionOfType(GetTypeOfFunction(f))
    Replace function f in l with f'
  end if

Algorithm 3 EvolveTrainingTasks

Require: Initial population of training tasks T_{init} (each task is a tuple $(\rho, \mathcal{E})$ where $\mathcal{E} = \{(I^{(i)}, O^{(i)}), \dots\}$), depth
Ensure: Updated task population T' (initialized with T_{init})
  $T \leftarrow T_{\text{init}}$
  $i \leftarrow 0$
  while $i < num_samples$ **do**
    if depth $= 1$ **then**
      $(\rho, \mathcal{E}) \leftarrow$ RandomSelectTask(T_{init}) {Select from initial tasks}
    else
      $(\rho, \mathcal{E}) \leftarrow$ RandomSelectTask(T) {Select from current tasks}
    end if
    $\rho' \leftarrow$ MutateProgram(ρ)
    $\mathcal{E}' \leftarrow \emptyset$ {Initialize mutated task demonstration examples}
    for each $(I^{(k)}, _) \in \mathcal{E}$ **do**
      $O'^{(k)} \leftarrow$ Execute($\rho', I^{(k)}$)
      $\mathcal{E}' \leftarrow \mathcal{E}' \cup \{(I^{(k)}, O'^{(k)})\}$
    end for
    if AreValidGrids(GetAllOutputs($\mathcal{E}'$)) **then**
      $T' \leftarrow T' \cup \{(\rho', \mathcal{E}')\}$ {Add new task to the population}
    end if
    $i \leftarrow i + 1$
  end while

Truncation We truncate our task demonstration tokens and program tokens such that these sequences fit in our predefined encoder and decoder context windows. For the task demonstration examples, we first order by grid size and divide the encoder context window into two equally sized sections. For task demonstration inputs, we first encode input grids to text as above and then we tokenize using the standard text tokenizer. We truncate these tokens at half the size of the encoder context window. We do the same for the task demonstration outputs and with the exception of also adding an end of sequence token. As a result, even though we aim to show the policy up to ten task demonstration examples, large grids will be cut-off. For programs, we tokenize directly using the standard text tokenizer and truncate at the decoder context window size.

A.4. ARC evaluation

Different works use different evaluation procedures to report performance on the ARC evaluation set. We describe two common evaluation settings in more detail below. Unless mentioned otherwise, we always use the first procedure, "ARC Eval Set".

ARC Eval Set This setup is intended as close as possible to the evaluation procedure described by Chollet (2019). Baselines Ferré (2021), Ainooson et al. (2023) follow this procedure, and it is our default setting as well.

The ARC eval set consists of 400 tasks, some of which contain multiple test examples. Common procedure is to report pass@3 performance, meaning the top 3 solutions are selected according to demonstration task performance. If there are ties, we favor the shorter program, under the assumption that shorter programs are more likely to generalize. We then run these programs on all test examples for the task. In some cases, there are multiple test examples per task. We call the task "solved" if all output grids are correct.

ARC Eval 412 This setup is designed to match Gendron et al. (2023). Instead of calling a task with multiple test examples solved if all test outputs are correct, distinct tasks are created - one per test example. This results in a set of 412 evaluation tasks with one test example each. Furthermore, Gendron et al. (2023) uses pass@1, rather than pass@3: only one solution per task is evaluated, and the task is considered solved if the output is correct.

B. Experiment details

B.1. Resources

Experiments were run for a maximum of 120 hours on a NVIDIA A100 80GB.

B.2. Hyperparameter tuning

Dataset The ARC benchmark does not contain a validation split. Hence, we use part of the ARC train split for validation during the hyperparameter tuning. In particular, this validation set is the search set that the sampling stage uses as described in 2.2. With this setup we avoid overfitting the hyperparameters to the ARC evaluation split.

We choose the split such that $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{valid}}$ contain roughly equally difficult programs by sampling based on program length: $\mathcal{D}_{\text{train}}$ contains 80% of 2-line programs, 80% of 3-line programs, and so on. This results in 311 examples in $\mathcal{D}_{\text{train}}$ and 89 examples in $\mathcal{D}_{\text{valid}}$.

Experiments on validation set In these experiments, we initialise our replay buffer with the 311 $\mathcal{D}_{\text{train}}$ examples, and our search set consists of the 89 $\mathcal{D}_{\text{valid}}$ examples. The aim of these experiments is to find optimal hyper-parameters for search and training. A list of our tuned hyperparameter values and their description is shown in Tab. 3

B.3. Hyperparamaters chosen on internal validation set

We optimized these parameters on our custom validation set before applying CodeIt to ARC eval.

CodeIt stage	Param	Value	Description
Sampling and Hindsight Relabeling	n_ρ	24	no. policy samples ρ per task per meta-iteration[1]
	n_m	$19,200$	no. mutated samples for augmented train set[1]
	τ	0.95	sampling temperature
	r_t	$10,000$	number of experiences sampled from augmented train set
	r_p	$90,000$	number of experiences sampled from buffer
Learning	n_ϵ	1	no. train epochs per meta-iteration
	lr	$5e-5$	learning rate

Table 3. Table of hyperparameters.

B.4. Domain Specific Language

We adopt the domain specific language (DSL) of Michael Hodel, made available on GitHub: https://github.com/michaelhodel/arc-dsl. This DSL was designed based on the training set: the (human) designer did not peek at the evaluation set. This is what allows us to run search on ARC eval here. Using a DSL designed for the eval tasks would be cheating, as we would benefit immensely from human insights captured in the primitives. On the other hand, it may mean that some ARC eval programs are not solvable with the current DSL.

The DSL is implemented in https://github.com/michaelhodel/arc-dsl/blob/main/dsl.py. It contains many basic grid manipulation operations, such as rotations (`rot90, rot180, rot270`), mirroring (`dmirror`, `hmirror`, `vmirror`), resizing (`downscale`, `upscale`), or concatenation (`hconcat`, `vconcat`). It also contains functions that perform counting, for example `numcolors` counts the number of colors occurring in an object or grid. For some ARC tasks, identifying the foreground objects and determining how these objects interact is an effective strategy for human test-takers. Therefore, some functions also apply to "objects", which are patches of the same color that stand out from the background. To extract these, the function `objects` returns the set of foreground objects, i.e. those that have a different color than the most common color, assumed to be the background. For a complete list of primitives and their description, we refer the reader to the aforementioned Github page.

Michael Hodel provides hand-designed solution programs for all training tasks in https://github.com/michaelhodel/arc-dsl/blob/main/solvers.py. Some programs are highly complex: for some of the more challenging ARC tasks, we see solutions consisting of up to 58 lines of code (`solve_b775ac94`). We use these 400 solution programs to kickstart CodeIt training.

C. ConceptARC

The ConceptARC dataset (Moskvichev et al., 2023) is similar to ARC with respect to the underlying task being a grid transformation problem, and the need to few-shot generalize from demonstration to test examples. However, tasks are divided into 16 concept groups; ConceptARC is moreover designed to be easier than ARC. We perform analysis on this additional benchmark, to highlight the type of tasks that CodeIt works well on.

We apply the same hyper-parameters as reported in Section B.3, with the exception of $n_p = 60$ which we rescale in proportion to the number of evaluation tasks in the ConceptARC dataset. The evaluation metric is also different: for ConceptARC, performance is evaluated separately on each test case per task. To ensure our results are comparable with existing work, we evaluate CodeIt by taking the model checkpoint at 100 meta-iterations, and perform pass@3 inference for temperature=0 (greedy decoding) and temperature=1 (simple sampling).

In Figure 8, we report CodeIt performance on ConceptARC with two evaluation temperatures against GPT-4 baselines from the literature, GPT-4 zero shot (Moskvichev et al., 2023) and GPT few shot refine (Mitchell et al., 2023). We observe that CodeIt outperforms GPT-4 with respect to the following concepts: extend to boundary, extract objects, filled not filled and same different. However, CodeIt substantially under-performs GPT-4 with respect to order, clean up and count. We note that while the DSL includes primitives for ordering and counting, it appears that CodeIt performs best on tasks related to object interactions, however, CodeIt performs worst on numerical or logic based tasks.

[1]Note that no. samples here refers to policy and mutation samples before filtering for syntactic correctness.

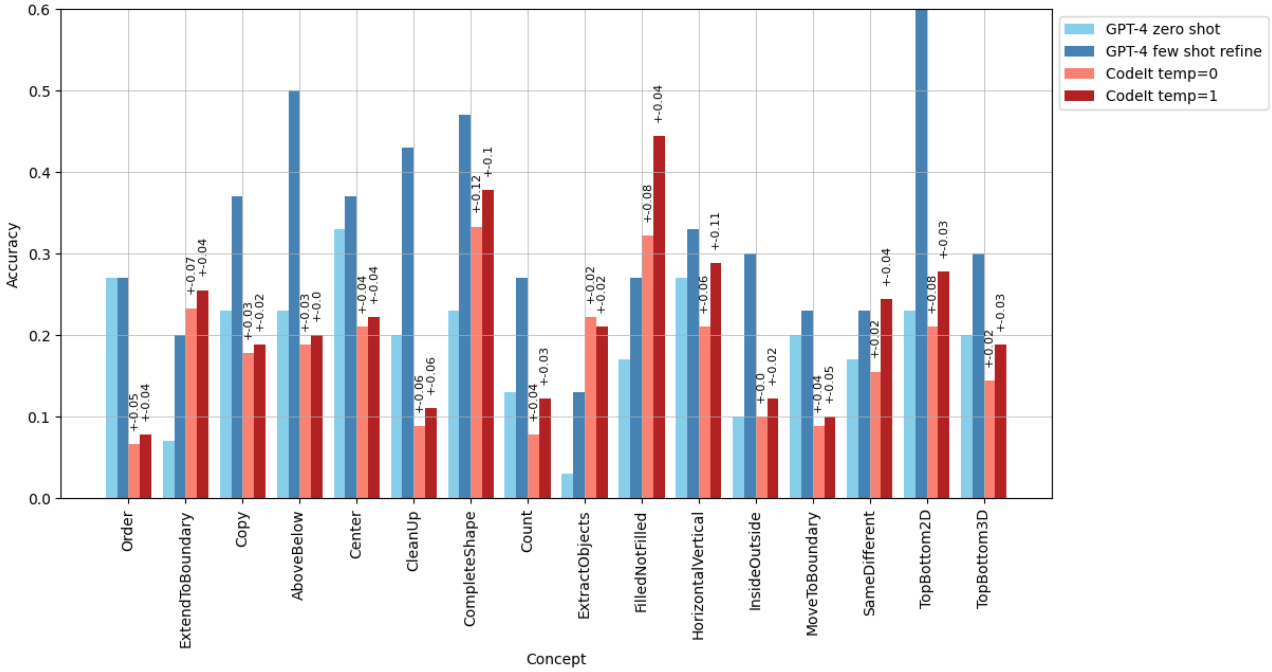


Figure 8. ConceptARC test accuracy CodeIt (three seeds) and GPT-4

D. Scaling Study

We perform a scaling study comparing CodeT5 model sizes and their effect on performance. Figure 9 shows a significant increase in performance increasing the model size from 60m to 220m and a smaller increase in performance from 220m to 770m. Note that for the 770m model, CodeIt was only run for 85 meta-iterations due to computation constraints.

It was our intention to add an even larger model to the scaling study (we opted for Mistral-7B). However, running CodeIt with this model took a lot longer (reaching 25 meta-iterations after 14 days) due to a the increased time needed for the sampling stage. We also found that replacing CodeIt's learning stage with LoRA-based finetuning (we carry out full-parameter finetuning with CodeT5) with a model of this size ultimately resulted in less than 9% solved tasks after 200k sampled programs. For comparison, the CodeT5 220M model solves close to 11%, and CodeT5 770M around 12% at that same point. A possible reason is that such large models are more prone to overfitting given the small size of the dataset. We do believe LoRA-based finetuning of a larger model can be made to work, but that it would require a proper search of LoRA hyperparameters, as well as benchmarking different LMs in order to find the best backbone model for this task; we deem this to be beyond the scope of the present work.

E. Program analysis

E.1. Failure Cases on Custom Validation Set

We analyze the properties of solved tasks on the custom validation set that we used for hyperparameter tuning. For this set, groundtruth programs are available, which allows us to analyze CodeIt's use of function primitives (the DSL operators). We ran CodeIt for 50 meta-iterations, and examined the program length and primitives of solved and unsolved tasks. Figure 11 visualises the results. We observe that CodeIt almost always solves the task if the solution program is short (10 lines or fewer). Further, CodeIt does not solve tasks of medium length (11 to 15 lines), but often learns to use their function primitives Finally, CodeIt never learns some primitives, especially those occurring in longer programs (16 lines or more). Perhaps unsurprisingly, CodeIt does not find a solution for tasks with a longer groundtruth program, as we can assume the tasks are more complex. Future work could address this by including information on the execution state during program generation, effectively breaking complex problems up into subproblems.

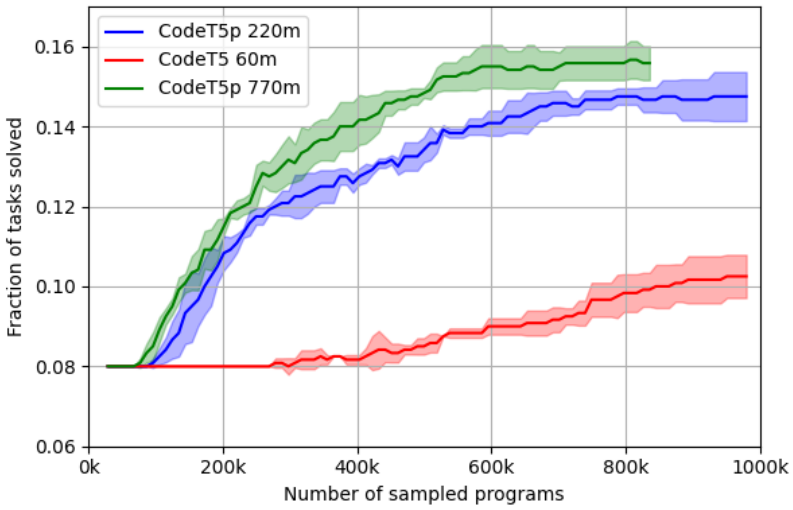


Figure 9. Cumulative performance as function of number of sampled programs for CodeIt with different model sizes

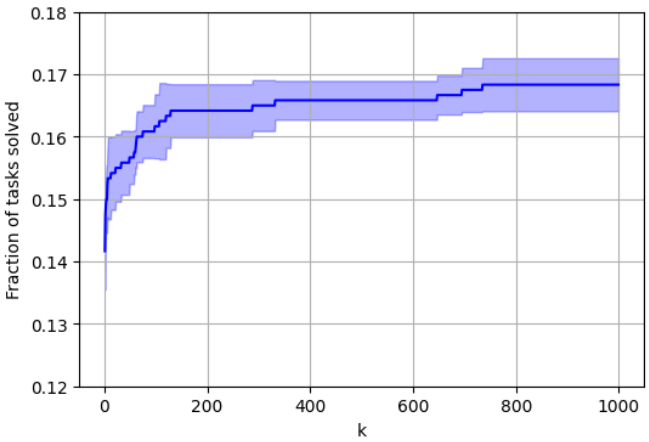


Figure 10. Pass@k curve for CodeIt over three seeds

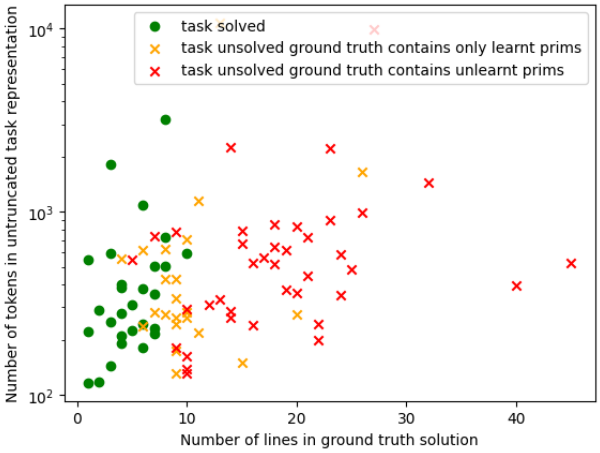


Figure 11. Failure cases on ARC custom validation set

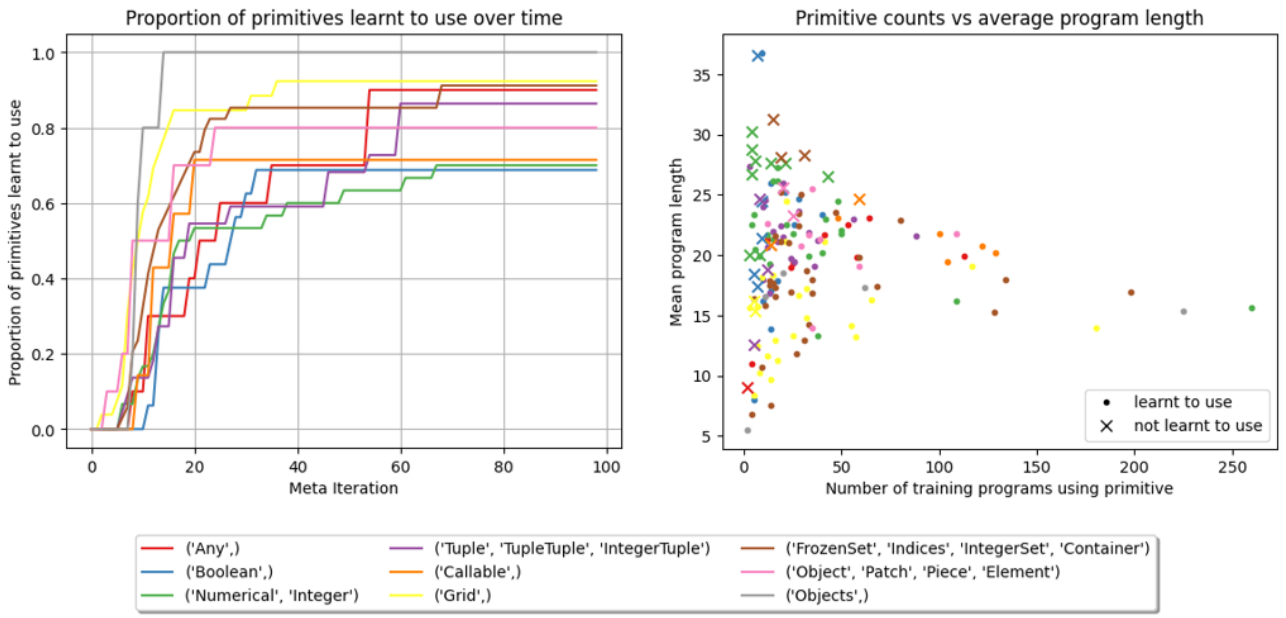


Figure 12. Primitives used in solutions for ARC evaluation set by output type. Left: proportion of primitives learnt over time. Right: scatter plot of counts vs mean program length.

Method	Number of tasks solved
CodeIt policy only	23
Mutation d_1 only	13
CodeIt policy $\cap$ Mutation d_1	29

Table 4. ARC evaluation tasks solved per method. The top group of two rows show how many tasks were solved by a method, but not by the other. The final row shows tasks solved by both methods.

E.2. DSL Learning

We investigate how CodeIt learns to use DSL primitives on the ARC evaluation set. We plot the use of primitives over training time, and a scatter plot indicating for each primitive whether it is used, in Figure 12. We find that some primitive types are learned quicker than others: for example, ones that return object representations are learned much quicker than ones that return numeric values. Moreover, primitives that are not learned typically occur in few training programs, or programs that are long, likely indicating it is more complex to learn to use these primitives.

E.3. CodeIt compared with mutation baselines

We compare the programs found using our mutation d_1 baseline and the best performing of the three CodeIt runs. Table 4 displays the number of ARC evaluation tasks uniquely solved by each method and the tasks which are solved by multiple methods. CodeIt's policy solves 52/400 tasks, 23 of which were not solved by the mutation baseline. In Figures 13 and 14, we select the shortest program that solves an evaluation task for CodeIt and our mutation d_1 baseline, computing the program length and task representation size. Note that CodeIt has an encoder context window size of 1024 and so any tasks which having representations of more than 1024 tokens have been truncated. Overall, CodeIt finds shorter programs as shown in 13. Further, for the same task, CodeIt more often finds shorter programs than our mutation d_1 baseline, as shown in 14 where each color represents a different task. Interestingly, CodeIt does solve some tasks with very large task representations, suggesting in some cases a truncated task representation provides sufficient information to solve the task.

In Table 5, we show a subset of solution programs for ARC eval tasks solved by both CodeIt and our mutation d_1 baseline. We select tasks where the shortest programs differ between the two methods. CodeIt programs appear more concise and use different primitives. Out of the 29 tasks that are solved by both methods, there are 24 shortest programs where the programs

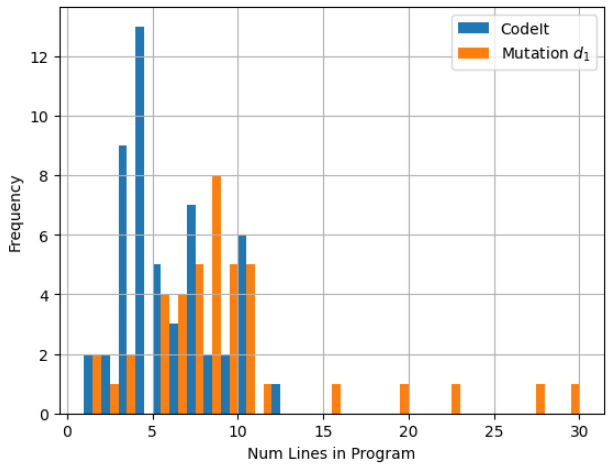


Figure 13. Histogram of number of lines for tasks where both CodeIt and Mutation produced solutions. CodeIt (in blue) produces shorter programs than the Mutation baseline (in orange).

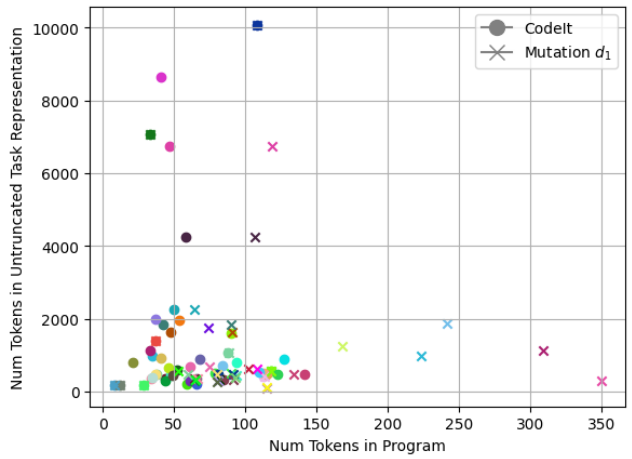


Figure 14. Number of task representation tokens vs number of program tokens. Colors represents the different tasks. We see no obvious correlation between task representation and program length.

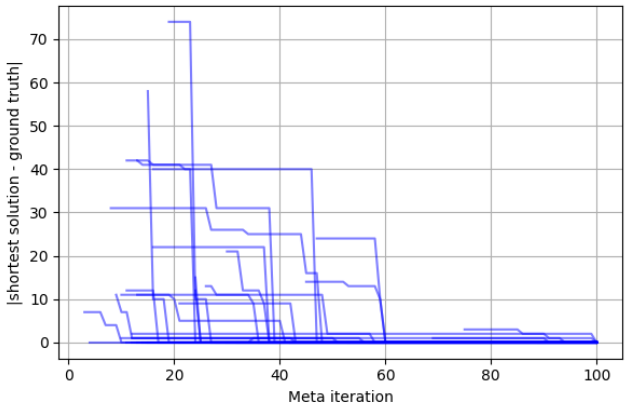


Figure 15. Difference in number of tokens between the shortest solution found per meta-iteration and shortest solution found by the final meta-iteration for best performing CodeIt run.

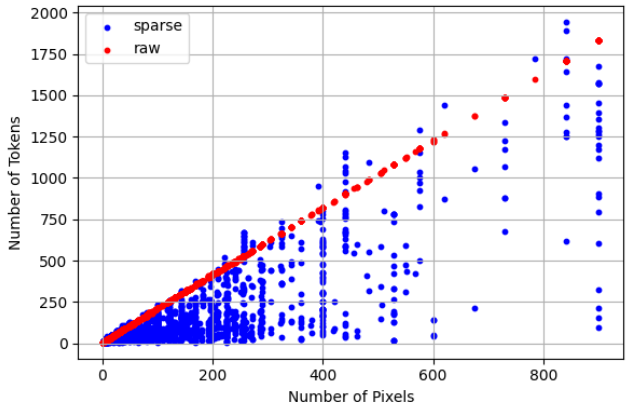


Figure 16. Grid size versus token count for the ARC training data. The sparse grid representation is typically shorter than the raw grid representation.

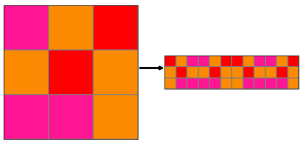
differ by method. CodeIt only produces a longer program in 1 out of these 24 cases. The Mutation baseline often includes redundant lines. In addition, for many programs, CodeIt produces a program that is qualitatively better: the solution is less complex, and contains fewer lines overall.

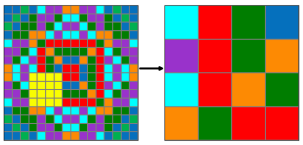
E.4. CodeIt over time

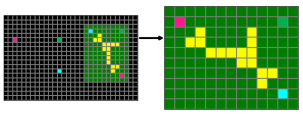
Since we do not have ground truth programs for the ARC evaluation set, we treat the shortest program found with demonstration performance and test performance equal to 1 for each task over all the meta-iterations as a proxy for the ground truth program. To examine how CodeIt solutions change over time, we take the subset of ARC evaluation tasks where the best performing CodeIt run finds such programs; this leaves us 45 tasks. We observe that once CodeIt finds a solution, CodeIt often continues to find both longer and shorter solutions in later meta-iterations. We pose that this gives the potential for program refinement, however, since the priority does not incorporate length, there is not explicit bias towards shorter solutions and so both longer and shorter solutions will be learned from. We observe that out of the 45 tasks, the best performing CodeIt run finds shorter solutions over time in 24 tasks as shown in Figure 15.

In Tables 6, we show a selection of examples where the best performing CodeIt run finds a longer solution in an earlier meta-iteration and shorter solution in a later meta-iteration.

CodeIt Policy	Mutation $d1$	Test Example

`x1 = vmirror(I)`
`x2 = hconcat(x1, I)`
`O = hconcat(x2, x2)`
	`x1 = vmirror(I)`
	`x2 = hconcat(x1, I)`
	`x3 = hmirror(x2)`
	`x4 = vconcat(x2, x3)`
	`x5 = hconcat(x3, x3)`
	`O = hmirror(x5)`
		

`x1 = compress(I)`
`x2 = ofcolor(I, THREE)`
`x3 = rot90(x1)`
`O = subgrid(x2, x3)`
	`x1 = hmirror(I)`
	`x2 = vmirror(I)`
	`x3 = ofcolor(I, THREE)`
	`x4 = subgrid(x3, x1)`
	`x5 = subgrid(x3, x2)`
	`x6 = palette(x4)`
	`x7 = contained(ONE, x6)`
	`O = branch(x7, x5, x4)`
		

`x1 = ofcolor(I, ONE)`
`x2 = subgrid(x1, I)`
`O = cmirror(x2)`
	`x1 = mostcolor(I)`
	`x2 = objects(I, T, F, T)`
	`x3 = replace(I, x1, THREE)`
	`x4 = argmax(x2, size)`
	`x5 = argmin(x2, size)`
	`x6 = position(x4, x5)`
	`x7 = first(x6)`
	`x8 = last(x6)`
	`x9 = subgrid(x4, x3)`
	`x10 = hline(x5)`
	`x11 = hmirror(x9)`
	`x12 = vmirror(x9)`
	`x13 = branch(x10, x11, x12)`
	`x14 = branch(x10, x7, ZERO)`
	`x15 = branch(x10, ZERO, x8)`
	`x16 = asobject(x13)`
	`x17 = matcher(first, THREE)`
	`x18 = compose(flip, x17)`
	`x19 = sfilter(x16, x18)`
	`x20 = ulcorner(x4)`
	`x21 = shape(x4)`
	`x22 = astuple(x14, x15)`
	`x23 = multiply(x21, x22)`
	`x24 = add(x20, x23)`
	`x25 = shift(x19, x24)`
	`x26 = rot270(x11)`
	`O = paint(x26, x25)`
		

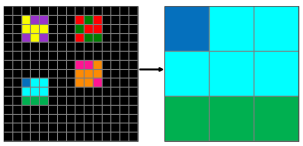
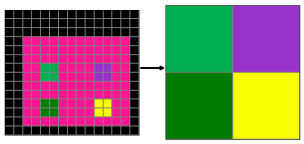
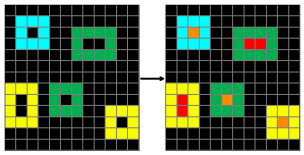
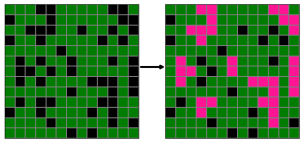
`x1 = objects(I, F, F, T)`
`x2 = argmax(x1, numcolors)`
`O = subgrid(x2, I)`
	`x1 = objects(I, F, F, T)`
	`x2 = leastcolor(I)`
	`x3 = rbind(colorcount, x2)`
	`x4 = argmax(x1, x3)`
	`O = subgrid(x4, I)`
		

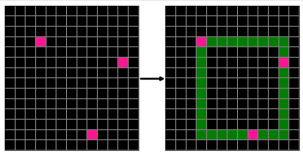
Table 5. Selection of shortest programs for ARC evaluation tasks solved by CodeIt policy (left) and the Mutation d_1 baseline (right) for which CodeIt program is shorter.

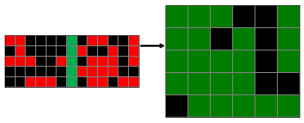
Early Shortest Solution	Later Shortest Solution	Test Example

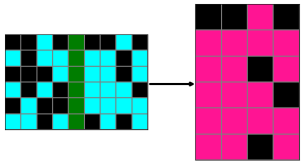
`x1 = ofcolor(I, EIGHT)`
`x2 = replace(I, EIGHT, ZERO)`
`x3 = compress(x2)`
`O = downscale(x3, TWO)`
	`x1 = replace(I, EIGHT, ZERO)`
	`x2 = compress(x1)`
	`O = downscale(x2, TWO)`
		

`x1 = objects(I, T, F, T)`
`x2 = apply(delta, x1)`
`x3 = mfilter(x2, square)`
`x4 = fill(I, FIVE, x3)`
`x5 = objects(x4, F, F, T)`
`x6 = mapply(delta, x5)`
`x7 = fill(x4, SEVEN, x6)`
`O = fill(x7, FIVE, x3)`
	`x1 = objects(I, T, F, T)`
	`x2 = apply(delta, x1)`
	`x3 = mfilter(x2, square)`
	`x4 = fill(I, FIVE, x3)`
	`x5 = objects(x4, F, F, T)`
	`x6 = mapply(delta, x5)`
	`O = fill(x4, SEVEN, x6)`
		

`x1 = objects(I, T, F, F)`
`x2 = colorfilter(x1, ZERO)`
`x3 = sizefilter(x2, ONE)`
`x4 = difference(x2, x3)`
`x5 = merge(x4)`
`O = fill(I, EIGHT, x5)`
	`x1 = objects(I, T, F, T)`
	`x2 = sizefilter(x1, ONE)`
	`x3 = difference(x1, x2)`
	`x4 = merge(x3)`
	`O = fill(I, EIGHT, x4)`
		

`x1 = vmirror(I)`
`x2 = fgpartition(I)`
`x3 = compose(outbox, inbox)`
`x4 = mapply(x3, x2)`
`O = underfill(I, ONE, x4)`
	`x1 = ofcolor(I, EIGHT)`
	`x2 = box(x1)`
	`O = underfill(I, ONE, x2)`
		

`x1 = lefthalf(I)`
`x2 = righthalf(I)`
`x3 = ofcolor(x1, ZERO)`
`x4 = ofcolor(x2, ZERO)`
`x5 = intersection(x3, x4)`
`x6 = shape(x1)`
`x7 = canvas(ONE, x6)`
`O = fill(x7, ZERO, x5)`
	`x1 = lefthalf(I)`
	`x2 = righthalf(I)`
	`x3 = cellwise(x1, x2, ONE)`
	`O = replace(x3, SEVEN, ONE)`
		

`x1 = lefthalf(I)`
`x2 = righthalf(I)`
`x3 = ofcolor(x1, FOUR)`
`x4 = ofcolor(x2, FOUR)`
`x5 = combine(x3, x4)`
`O = fill(x1, EIGHT, x5)`
	`x1 = lefthalf(I)`
	`x2 = righthalf(I)`
	`x3 = cellwise(x1, x2, FOUR)`
	`O = replace(x3, FOUR, EIGHT)`
		

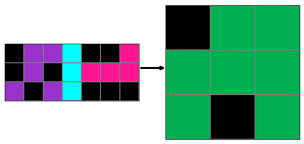
`x1 = lefthalf(I)`
`x2 = righthalf(I)`
`x3 = ofcolor(x1, ZERO)`
`x4 = ofcolor(x2, ZERO)`
`x5 = intersection(x3, x4)`
`x6 = shape(x1)`
`x7 = canvas(TWO, x6)`
`O = fill(x7, ZERO, x5)`
	`x1 = vmirror(I)`
	`x2 = lefthalf(I)`
	`x3 = righthalf(I)`
	`x4 = cellwise(x2, x3, TWO)`
	`O = replace(x4, EIGHT, TWO)`
		

Table 6. Selection of shortest solutions for ARC evaluation tasks solved by CodeIt policy where shorter solutions are found over time.

F. ARC competitions

Competition	Winner	Method	Hidden Test Perf.
Kaggle 2020	Icecuber (2020)	Search in eval set DSL*	21%
Kaggle 2020 late	Multiple (Kaggle Leaderboard, 2020)	Ensemble previous entries*	30%
ARCathon 2022	Hodel (2023)	Search in CodeIt DSL	6%
ARCathon 2023	Multiple (ARCathon Leaderboard, 2023)	Unknown	30%

Table 7. Performance on Hidden Test Set for Various ARC Competition Winners. *Method conditions on ARC evaluation set.