

Opening the AI black box: program synthesis via mechanistic interpretability

Eric J. Michaud^{*12} Isaac Liao^{*1} Vedang Lad^{*1} Ziming Liu^{*12} Anish Mudide¹ Chloe Loughridge³
Zifan Carl Guo¹ Tara Rezaei Kheirkhah¹ Mateja Vukelić¹ Max Tegmark¹²

Abstract

We present MIPS, a novel method for program synthesis based on automated mechanistic interpretability of neural networks trained to perform the desired task, auto-distilling the learned algorithm into Python code. We test MIPS on a benchmark of 62 algorithmic tasks that can be learned by an RNN and find it highly complementary to GPT-4: MIPS solves 32 of them, including 13 that are not solved by GPT-4 (which also solves 30). MIPS uses an integer autoencoder to convert the RNN into a finite state machine, then applies Boolean or integer symbolic regression to capture the learned algorithm. As opposed to large language models, this program synthesis technique makes no use of (and is therefore not limited by) human training data such as algorithms and code from GitHub. We discuss opportunities and challenges for scaling up this approach to make machine-learned models more interpretable and trustworthy.

1. Introduction

Machine-learned algorithms now outperform traditional human-discovered algorithms on many tasks, from translation to general-purpose reasoning. These learned algorithms tend to be black-box neural networks, and we typically lack a full understanding of how they work. This has motivated the growing field of *mechanistic interpretability*, aiming to assess and improve their trustworthiness. Major progress has been made in interpreting and understanding smaller models, but this work has involved human effort, which raises questions about whether it can scale to larger models. This makes it timely to investigate whether mechanistic

interpretability can be fully automated (Tegmark & Omonhundro, 2023).

The goal of the present paper is to take a modest first step in this direction by presenting and testing MIPS (**M**echanistic **I**nterpretability-based **P**rogram **S**ynthesis), a fully automated method that can distill simple learned algorithms from neural networks into Python code. The rest of this paper is organized as follows. After reviewing prior work in Section II, we present our method in Section III, test it on a benchmark in Section IV and summarize our conclusions in Section V.

2. Related Work

Program synthesis is a venerable field dating back to Alonzo Church in 1957; Zhou & Ding (2023) and Odena et al. (2020) provide recent reviews of the field. Large language Models (LLMs) have become increasingly good at writing code based on verbal problem descriptions or auto-complete. We instead study the common alternative problem setting known as “programming by example” (PBE), where the desired program is specified by giving examples of input-output pairs (Wu et al., 2023). The aforementioned papers review a wide variety of program synthesis methods, many of which involve some form of search over a space of possible programs. LLMs that synthesize code directly have recently become quite competitive with such search-based approaches (Sobania et al., 2022). Our work provides an alternative search-free approach where the program learning happens during neural network training rather than execution.

Our work builds on the recent progress in *mechanistic interpretability* (MI) of neural networks (Olah et al., 2020; Cammarata et al., 2020; Wang et al., 2022; Olsson et al., 2022). Much MI work has tried to understand how neural networks represent various types of information, *e.g.*, geographic information (Goh et al., 2021; Gurnee & Tegmark, 2023), truth (Burns et al., 2022; Marks & Tegmark, 2023) and the state of board games (McGrath et al., 2022; Toshniwal et al., 2022; Li et al., 2022). Another major MI thrust has been to understand how neural networks perform algorithmic tasks,

^{*}Equal contribution, alphabetical order ¹Massachusetts Institute of Technology, Cambridge, MA, USA ²Institute for Artificial Intelligence and Fundamental Interactions ³Harvard University, Cambridge, MA, USA. Correspondence to: Max Tegmark <tegmark@mit.edu>.

e.g., modular arithmetic (Nanda et al., 2023; Liu et al., 2022; Zhong et al., 2023; Quirke et al., 2023), greater-than (Hanna et al., 2023), and greatest-common-divisor (Charton, 2023).

Whereas Lindner et al. (2023) automatically convert traditional code into a neural network, we aim to do the opposite. Other recent efforts to automate MI include identifying a sparse subgraph of the network whose units are causally relevant to a behavior of interest (Conmy et al., 2023; Syed et al., 2023) and using LLMs to label internal components of neural networks, for instance neurons (Bills et al., 2023) and features discovered by sparse autoencoders (Cunningham et al., 2023; Bricken et al., 2023).

3. MIPS, our program synthesis algorithm

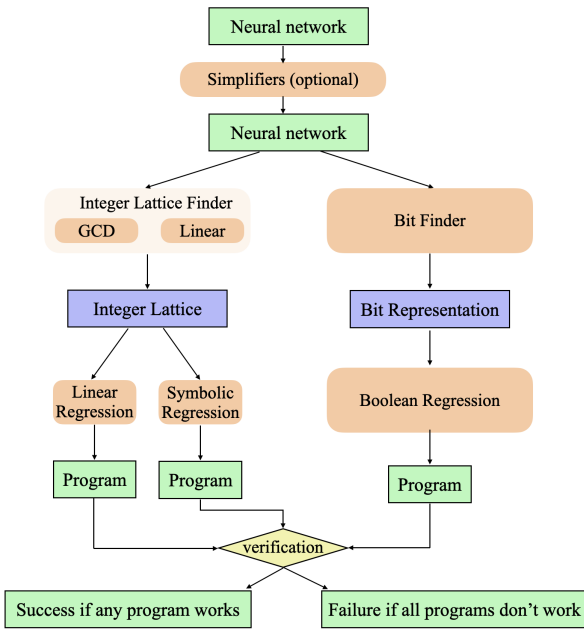


Figure 1. The pipeline of our program synthesis method. MIPS relies on discovering integer representations and bit representations of hidden states, which enable regression methods to figure out the exact symbolic relations between input, hidden, and output states.

As summarized in Figure 1, MIPS involves the following key steps.

1. Neural network training
2. Neural network simplification
3. Finite state machine extraction
4. Symbolic regression

Step 1 is to train a black-box neural network to learn an algorithm that performs the desired task. In this paper, we

use a recurrent neural network (RNN) of the general form

$$\mathbf{h}_i = f(\mathbf{h}_{i-1}, \mathbf{x}_i), \quad (1)$$

$$\mathbf{y}_i = g(\mathbf{h}_i), \quad (2)$$

that maps input vectors \mathbf{x}_i into output vectors \mathbf{y}_i via hidden states \mathbf{h}_i . The RNN is defined by the two functions f and g , which are implemented as feed-forward neural networks (MLPs) to allow more model expressivity than for a vanilla RNN. The techniques described below can also be applied to more general neural network architectures.

Step 2 attempts to automatically simplify the learned neural network without reducing its accuracy. Steps 3 and 4 automatically distill this simplified learned algorithm into Python code. When the training data is discrete (consisting of say text tokens, integers, or pixel colors), the neural network will be a *finite state machine*: the activation vectors for each of its neuron layers define finite sets and the entire working of the network can be defined by look-up tables specifying the update rules for each layer. For our RNN, this means that the space of hidden states \mathbf{h} is discrete, so that the functions f and g can be defined by lookup tables. As we will see below, the number of hidden states that MIPS needs to keep track of can often be greatly reduced by clustering them, corresponding to learned representations. After this, the geometry of the cluster centers in the hidden space often reveals that they form either an incomplete multidimensional lattice whose points represent integer tuples, or a set whose cardinality is a power of two, whose points represent Boolean tuples. In both of these cases, the aforementioned lookup tables simply specify integer or Boolean functions, which MIPS attempts to discover via symbolic regression. Below we present an *integer autoencoder* and a *Boolean autoencoder* to discover such integer/Boolean representations from arbitrary point sets.

We will now describe each of the three steps of MIPS in greater detail.

3.1. AutoML optimizing for simplicity

We wish to find the *simplest* RNN that can learn our task, to facilitate subsequent discovery of the algorithm that it has learned. We therefore implement an AutoML-style neural architecture search that tries to minimize network size while achieving perfect test accuracy. This search space is defined by a vector \mathbf{p} of five main architecture hyperparameters: the five integers $\mathbf{p} = (n, w_f, d_f, w_g, d_g)$ corresponding to the dimensionality of hidden state \mathbf{h} , the width and depth of the f -network, and the width and depth of the g -network, respectively. Both the f - and g -networks have a linear final layer and ReLU activation functions for all previous layers. The hidden state \mathbf{h}_0 is initialized to zero.

To define the parameter search space, we define ranges for each parameter. For all tasks, we use $n \in$

$\{1, 2, \dots, 128\}$, $w_f \in \{1, 2, \dots, 256\}$, $d_f \in \{1, 2, 3\}$, $w_g \in \{1, 2, \dots, 256\}$ and $d_g \in \{1, 2, 3\}$, so the total search space consists of $128 \times 256 \times 3 \times 256 \times 3 = 75,497,472$ hyperparameter vectors \mathbf{p}_i . We order this search space by imposing a strict ordering on the importance of minimizing each hyperparameter – lower d_g is strictly more important than lower d_f , which is strictly more important than lower n , which is strictly more important than lower w_g , which is strictly more important than lower w_f . We aim to find the hyperparameter vector (integer 5-tuple) p_i in the search space which has lowest rank i under this ordering.

We search the space in the following simple manner. We first start at index $i = 65,536$, which corresponds to parameters $(1, 1, 2, 1, 1)$. For each parameter tuple, we train networks using 5 different seeds. We use the loss function $\ell(x, y) = \frac{1}{2} \log[1 + (x - y)^2]$, finding that it led to more stable training than using vanilla MSE loss. We train for either 10,000 or 20,000 steps, depending on the task, using the Adam optimizer, a learning rate of 10^{-3} , and batch size 4096. The test accuracy is evaluated with a batch of 65536 samples. If no networks achieve 100% test accuracy (on any test batch), we increase i by $2^{1/4}$. We proceed in this manner until we find a network where one of the seeds achieves perfect test accuracy or until the full range is exhausted. If we find a working network on this upwards sweep, we then perform a binary search using the interval halving method, starting from the successful i , to find the lowest i where at least one seed achieves perfect test accuracy.

3.2. Auto-simplification

After finding a minimal neural network architecture that can solve a task, the resulting neural network weights typically seem random and un-interpretable. This is because there exist symmetry transformations of the weights that leave the overall input-output behavior of the neural network unchanged. The random initialization of the network has therefore caused random symmetry transformations to be applied to the weights. In other words, the learned network belongs to an equivalence class of neural networks with identical behavior and performance, corresponding to a submanifold of the parameter space. We exploit these symmetry transformations to simplify the neural network into a *normal form*, which in a sense is the simplest member of its equivalence class. Conversion of objects into a normal/standard form is a common concept in mathematics and physics (for example, conjunctive normal form, wavefunction normalization, reduced row echelon form, and gauge fixing).

Two of our simplification strategies below exploit a symmetry of the RNN hidden state space \mathbf{h} . We can always write the MLP g in the form $g(\mathbf{h}) = G(\mathbf{U}\mathbf{h} + \mathbf{c})$ for some function G . So if f is affine, *i.e.*, of the form

$f(\mathbf{h}, \mathbf{x}) = \mathbf{W}\mathbf{h} + \mathbf{V}\mathbf{x} + \mathbf{b}$, then the symmetry transformation

$\mathbf{W}' \equiv \mathbf{A}\mathbf{W}\mathbf{A}^{-1}$, $\mathbf{V}' = \mathbf{A}\mathbf{V}$, $\mathbf{U}' = \mathbf{U}\mathbf{A}^{-1}$, $\mathbf{h}' \equiv \mathbf{A}\mathbf{h}$, $\mathbf{b}' = \mathbf{A}\mathbf{b}$ keeps the RNN in the same form:

$$\begin{aligned} \mathbf{h}'_i &= \mathbf{A}\mathbf{h}_i = \mathbf{A}\mathbf{W}\mathbf{A}^{-1}\mathbf{A}\mathbf{h}_{i-1} + \mathbf{A}\mathbf{V}\mathbf{x}_i + \mathbf{A}\mathbf{b} \\ &= \mathbf{W}'^{-1}\mathbf{h}'_{i-1} + \mathbf{V}'\mathbf{x}_i + \mathbf{b}', \end{aligned} \quad (3)$$

$$\begin{aligned} \mathbf{y}_i &= G(\mathbf{U}\mathbf{h}_i + \mathbf{c}) = G(\mathbf{U}\mathbf{A}^{-1}\mathbf{h}'_i + \mathbf{c}) \\ &= G(\mathbf{U}'\mathbf{h}'_i + \mathbf{c}). \end{aligned} \quad (4)$$

We think of neural networks as nails, which can be hit by various auto-normalization hammers. Each hammer is an algorithm that applies transformations to the weights to remove degrees of freedom caused by extra symmetries or cleans the neural network up in some other way. In this section, we describe five normalizers we use to simplify our trained networks, termed “Whitening”, “Jordan normal form”, “Toeplitz”, “De-bias”, and “Quantization”. For every neural network, we always apply this sequence of normalizers in that specific order, for consistency. We describe them below and provide additional details about them in the Appendix D.

1. **Whitening:** Just as we normalize input data to use for training neural networks, we would like activations in the hidden state space \mathbf{h}_i to be normalized. To ensure normalization in all directions, we feed the training dataset into the RNN, collect all the hidden states, compute the uncentered covariance matrix \mathbf{C} , and then apply a whitening transform $\mathbf{h} \mapsto \mathbf{C}^{-1/2}\mathbf{h}$ to the hidden state space so that its new covariance becomes the identity matrix. This operation exists purely to provide better numerical stability to the next step.
2. **Jordan normal form:** When the function g is affine, we can apply the aforementioned symmetry transformation to try to diagonalize \mathbf{W} , so that none of the hidden state dimensions interact with one another. Unfortunately, not all matrices \mathbf{W} can be diagonalized, so we use a generalized alternative: the Jordan normal form, which allows elements of the superdiagonal to be either zero or one. To eliminate complex numbers, we also apply 2×2 unitary transformations to eigenvectors corresponding to conjugate pairs of complex eigenvalues afterward. The aforementioned whitening is now ruined, but it helped make the Jordan normal form calculation more numerically stable.
3. **Toeplitz:** Once \mathbf{W} is in Jordan normal form, we divide it up into Jordan blocks and apply upper-triangular Toeplitz transformations to the dimensions belonging to each Jordan block. There is now an additional symmetry, corresponding to multiplying each Jordan block by an upper triangular Toeplitz matrix, and we exploit

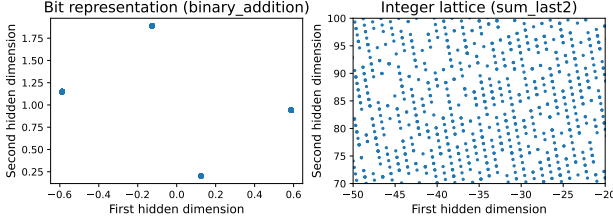


Figure 2. These hidden structures can be turned into discrete representations. Left: the hidden states for the bitstring addition task are seen to form four clusters, corresponding to 2 bits: the output bit and the carry bit. Right: the hidden states for the Sum_Last2 task are seen to form clusters on a 2D lattice corresponding to two integers.

the Toeplitz matrix that maximally simplifies the aforementioned \mathbf{V} -matrix.

4. **De-bias:** Sometimes \mathbf{W} is not full rank, and \mathbf{b} has a component in the direction of the nullspace. In this case, the component can be removed, and the bias \mathbf{c} can be adjusted to compensate.
5. **Quantization:** After applying all the previous normalizers, many of the weights may have become close to integers, but not exactly due to machine precision and training imperfections. Sometimes, depending on the task, all of the weights can become integers. We therefore round any weights that are within $\epsilon \equiv 0.01$ of an integer to that integer.

3.3. Boolean and integer autoencoders

As mentioned, our goal is to convert a trained recurrent neural network (RNN) into a maximally simple (Python) program that produces equivalent input-output behavior. This means that if the RNN has 100% accuracy for a given dataset, so should the program, with the added benefit of being more interpretable, precise, and verifiable.

Once trained/written, the greatest difference between a neural network and a program implementing the same finite state machine is that the former is fuzzy and continuous, while the latter is precise and discrete. To convert a neural network to a program, some discretization (“defuzzification”) process is needed to extract precise information from seemingly noisy representations. Fortunately, mechanistic interpretability research has shown that neural networks tend to learn meaningful, structured knowledge representations for algorithmic tasks (Liu et al., 2022; Nanda et al., 2023). Previous interpretability efforts typically involved case-by-case manual inspection, and only gained algorithmic understanding at the level of pseudocode at best. We tackle this more ambitious question: can we create an automated method that distills the learned representation and associated algorithms into an equivalent (Python) program?

Since the tasks in our benchmark involve bits and integers, which are already discrete, the only non-discrete parts in a recurrent neural network are its hidden representations. Here we show two cases when hidden states can be discretized: they are (1) a bit representation or (2) a (typically incomplete) integer lattice. Generalizing to the mixed case of bits and integers is straightforward. Figure 2 shows all hidden state activation vectors \mathbf{h}_i for all steps with all training examples for two of our tasks. The left panel shows that the 10^4 points \mathbf{h}_i form $2^2 = 4$ tight clusters, which we interpret as representing 2 bits. The right panel reveals that the points \mathbf{h}_i form an incomplete 2D lattice that we interpret as secretly representing a pair of integers.

Bit representations

The hidden states for the 2 bits in Figure 2 are seen to form a parallelogram. More generally, we find that hidden states encode b bits as 2^b clusters, which in some cases form b -dimensional parallelograms and in other cases look more random. Our algorithm tries all $(2^b)!$ possible assignments of the 2^b clusters to bitstrings of length b and selects the assignment that minimizes the length of the resulting Python program.

Integer lattice

As seen in Figure 2, the learned representation of an integer lattice tends to be both non-square (deformed by a random affine transformation) and sparse (since not all integer tuples occur during training). We thus face the following problem: given (possibly sparse) samples of points \mathbf{h}_i from an n -dimensional lattice, how can we reconstruct the integer tuple each lattice point represents? We call the solution an *integer autoencoder* since it compresses any point set into a set of integer tuples from which the original points can be at least approximately recovered as $\mathbf{h}_i = \mathbf{A}\mathbf{k}_i + \mathbf{b}$, where \mathbf{A} is a matrix and \mathbf{b} is a vector that defines the affine transformation and a set of integer vectors \mathbf{k}_i .

In the Appendix A, we present a solution that we call the *GCD lattice finder*. For the special case $n = 1$, its core idea is to compute the greatest common denominator of pairwise separations: for example, for the points $\{1.7, 3.2, 6.2, 7.7\}$, all point separations are divisible by $A = 1.5$, from which one infers that $b = 0.2$ and the lattice can be rewritten as $1.5 \times \{1, 2, 4, 5\} + 0.2$. For multidimensional lattices, our algorithm uses the GCD of ratios of generalized cell volumes to infer the directions and lengths of the lattice vectors that form the columns of \mathbf{A} .

For the special case where the MLP defining the function f is affine or can be accurately approximated as affine, we use a simpler method we term the *Linear lattice finder*, also described in Appendix B. Here the idea is to exploit that the lattice is simply an affine transformation of a regular integer

lattice (the input data), so we can simply “read off” the desired lattice basis vectors from this affine transformation.

Symbolic regression

Once the hidden states h_i have been successfully mapped to Boolean or integer tuples as described above, the functions f and g that specify the learned RNN can be re-expressed as lookup tables, showing their Boolean/integer output tuple for each Boolean/integer input tuple. All that remains is now *symbolic regression*, i.e., discovering the simplest possible symbolic formulae that define f and g .

Boolean regression: In the case where a function maps bits to a bit, our algorithm determines the following set of correct Boolean formulae and then returns the shortest one. The first candidate formula is the function written in disjunctive normal form, which is always possible. If the Boolean function is *symmetric*, i.e., invariant under all permutations of its arguments, then we also write it as an integer function of its bit sum.

Integer regression: In the case when a function maps integers to an integer, we try the following two methods:

1. If the function is linear, then we perform simple linear regression, round the resulting coefficients to integers, and simplify, e.g., multiplications by 0 and 1.
2. Otherwise, we use the brute-force symbolic solver from *AI Feynman* (Udrescu et al., 2020), including the 6 unary operations $\{>, <, \sim, H, D, A\}$ and 4 binary operations $\{+, -, *, \%\}$ whose meanings are explained in Appendix C, then convert the simplest discovered formula into Python format.

Once symbolic formulas have been separately discovered for each component of the vector-valued functions f and g , we insert them into a template Python program that implements the basic loop over inputs that are inherent in an RNN. We present examples of our auto-generated programs in Figures 3 and 4 and in Appendix F.

4. Results

We will now test the program synthesis abilities of our MIPS algorithm on a benchmark of algorithmic tasks specified by numerical examples. For comparison, we try the same benchmark on GPT-4 Turbo, which is currently (as of January 2024) described by OpenAI as their latest generation model, with a 128k context window and more capable than the original GPT-4.

4.1. Benchmark

Our benchmark consists of the 62 algorithmic tasks listed in Table 1. They each map one or two integer lists of length

10 or 20 into a new integer list. We refer to integers whose range is limited to $\{0, 1\}$ as bits. We generated this task list manually, attempting to produce a collection of diverse tasks that would in principle be solvable by an RNN. We also focused on tasks whose known algorithms involved majority, minimum, maximum, and absolute value functions because we believed they would be more easily learnable than other nonlinear algorithms due to our choice of the ReLU activation for our RNNs. The benchmark training data and project code is available at <https://github.com/ejmichaud/neural-verification>. The tasks are described in Table 1, with additional details in Appendix E.

Since the focus of our paper is not on whether RNNs can learn algorithms, but on whether learned algorithms can be auto-extracted into Python, we discarded from our benchmark any generated tasks on which our RNN-training failed to achieve 100% accuracy.

Our benchmark can never show that MIPS outperforms any large language model (LLM). Because LLMs are typically trained on GitHub, many LLMs can produce Python code for complicated programming tasks that fall outside of the class we study. Instead, the question that our MIPS-LLM comparison addresses is whether MIPS complements LLMs by being able to solve some tasks where an LLM fails.

4.2. Evaluation

For both our method and GPT-4 Turbo, a task is considered solved if and only if a Python program is produced that solves the task with 100% accuracy. GPT-4 Turbo is prompted using the “chain-of-thought” approach described below and illustrated in Figure 5.

For a given task, the LLM receives two lists of length 10 sourced from the respective RNN training set. The model is instructed to generate a formula that transforms the elements of list “x” (features) into the elements of list “y” (labels). Subsequently, the model is instructed to translate this formula into Python code. The model is specifically asked to use elements of the aforementioned lists as a test case and print “Success” or “Failure” if the generated function achieves full accuracy on the test case. An external program extracts a fenced markdown codeblock from the output, which is saved to a separate file and executed to determine if it successfully completes the task. To improve the chance of success, this GPT-4 Turbo prompting process is repeated three times, requiring only at least one of them to succeed. We run GPT using default temperature settings.

Table 1. Benchmark results. For tasks with note “see text”, please refer to Appendix E

Task #	Input Strings	Element Type	Task Description	Task Name	Solved by GPT-4?	Solved by MIPS?
1	2	bit	Binary addition of two bit strings	Binary_Addition	0	1
2	2	int	Ternary addition of two digit strings	Base_3_Addition	0	0
3	2	int	Base 4 addition of two digit strings	Base_4_Addition	0	0
4	2	int	Base 5 addition of two digit strings	Base_5_Addition	0	0
5	2	int	Base 6 addition of two digit strings	Base_6_Addition	1	0
6	2	int	Base 7 addition of two digit strings	Base_7_Addition	0	0
7	2	bit	Bitwise XOR	Bitwise_Xor	1	1
8	2	bit	Bitwise OR	Bitwise_Or	1	1
9	2	bit	Bitwise AND	Bitwise_And	1	1
10	1	bit	Bitwise NOT	Bitwise_Not	1	1
11	1	bit	Parity of last 2 bits	Parity_Last2	1	1
12	1	bit	Parity of last 3 bits	Parity_Last3	0	1
13	1	bit	Parity of last 4 bits	Parity_Last4	0	0
14	1	bit	Parity of all bits seen so far	Parity_All	0	1
15	1	bit	Parity of number of zeros seen so far	Parity_Zeros	0	1
16	1	int	Cumulative number of even numbers	Evens_Counter	0	0
17	1	int	Cumulative sum	Sum_All	1	1
18	1	int	Sum of last 2 numbers	Sum_Last2	0	1
19	1	int	Sum of last 3 numbers	Sum_Last3	0	1
20	1	int	Sum of last 4 numbers	Sum_Last4	1	1
21	1	int	Sum of last 5 numbers	Sum_Last5	1	1
22	1	int	sum of last 6 numbers	Sum_Last6	1	1
23	1	int	Sum of last 7 numbers	Sum_Last7	1	1
24	1	int	Current number	Current_Number	1	1
25	1	int	Number 1 step back	Prev1	1	1
26	1	int	Number 2 steps back	Prev2	1	1
27	1	int	Number 3 steps back	Prev3	1	1
28	1	int	Number 4 steps back	Prev4	1	1
29	1	int	Number 5 steps back	Prev5	1	1
30	1	int	1 if last two numbers are equal	Previous_Equals_Current	0	1
31	1	int	current — previous	Diff_Last2	0	1
32	1	int	current — previous	Abs.Diff	0	1
33	1	int	current	Abs.Current	1	1
34	1	int	current — previous	Diff_Abs.Values	1	0
35	1	int	Minimum of numbers seen so far	Min.Seen	1	0
36	1	int	Maximum of integers seen so far	Max.Seen	1	0
37	1	int	integer in 0-1 with highest frequency	Majority_0.1	1	0
38	1	int	Integer in 0-2 with highest frequency	Majority_0.2	0	0
39	1	int	Integer in 0-3 with highest frequency	Majority_0.3	0	0
40	1	int	1 if even, otherwise 0	Evens_Detector	1	0
41	1	int	1 if perfect square, otherwise 0	Perfect_Square_Detector	0	0
42	1	bit	1 if bit string seen so far is a palindrome	Bit_Palindrome	1	0
43	1	bit	1 if parentheses balanced so far, else 0	Balanced_Parenthesis	0	0
44	1	bit	Number of bits seen so far mod 2	Parity_Bits_Mod2	1	0
45	1	bit	1 if last 3 bits alternate	Alternating_Last3	0	0
46	1	bit	1 if last 4 bits alternate	Alternating_Last4	1	0
47	1	bit	bit shift to right (same as prev1)	Bit_Shift_Right	1	1
48	2	bit	Cumulative dot product of bits mod 2	Bit_Dot_Prod_Mod2	0	1
49	1	bit	Binary division by 3 (see text)	Div_3	1	0
50	1	bit	Binary division by 5 (see text)	Div_5	0	0
51	1	bit	Binary division by 7 (see text)	Div_7	0	0
52	1	int	Cumulative addition modulo 3	Add_Mod_3	1	1
53	1	int	Cumulative addition modulo 4	Add_Mod_4	0	0
54	1	int	Cumulative addition modulo 5	Add_Mod_5	0	0
55	1	int	Cumulative addition modulo 6	Add_Mod_6	0	0
56	1	int	Cumulative addition modulo 7	Add_Mod_7	0	0
57	1	int	Cumulative addition modulo 8	Add_Mod_8	0	0
58	1	int	1D dithering, 4-bit to 1-bit (see text)	Dithering	1	0
59	1	int	Newton’s of - freebody (integer input)	Newton.Freebody	0	1
60	1	int	Newton’s law of gravity (see text)	Newton.Gravity	0	1
61	1	int	Newton’s law w. spring (see text)	Newton.Spring	0	1
62	2	int	Newton’s law w. magnetic field (see text)	Newton.Magnetic	0	0
Total solved					30	32

```

1
2 def f(s,t):
3     a = 0;b = 0;
4     ys = []
5     for i in range(10):
6         c = s[i]; d = t[i];
7         next_a = b ^ c ^ d
8         next_b = b+c+d>1
9         a = next_a;b = next_b;
10        y = a
11        ys.append(y)
12    return ys
    
```

Figure 3. The generated program for the addition of two binary numbers represented as bit sequences. Note that MIPS rediscovers the “ripple adder”, where the variable b above is the carry bit.

4.3. Performance

As seen in Table 1, MIPS is highly complementary to GPT-4 Turbo: MIPS solves 32 of our tasks, including 13 that are not solved by ChatGPT-4 (which solves 30).

The AutoML process of Section 3.1 discovers networks of varying task-dependent shape and size. Table 2 shows the parameters \mathbf{p} discovered for each task. Across our 62 tasks, 16 tasks could be solved by a network with hidden dimension $n = 1$, and the largest n required was 81. For many tasks, there was an interpretable meaning to the shape of the smallest network we discovered. For instance, on tasks where the output is the element occurring k steps earlier in the list, we found $n = k + 1$, since the current element and the previous k elements must be stored for later recall.

We found two main failure modes for MIPS:

1. Noise and non-linearity. The latent space is still close to being a finite state machine, but the non-linearity and/or noise present in an RNN is so dominant that the integer autoencoder fails, *e.g.*, for *Diff_Abs_Values*. Humans can stare at the lookup table and regress the symbolic function with their brains, but since the lookup table is not perfect, *i.e.*, has the wrong integer in a few examples, MIPS fails to symbolically regress the function. This can probably be mitigated by learning and generalizing from a training subset with a smaller dynamic range.
2. Continuous computation. A key assumption of MIPS is that RNNs are finite-state machines. However, RNNs can also use continuous variables to represent information — the *Majority_0_X* tasks fail for this reason. This can probably be mitigated by identifying and implementing floating-point variables.

Figure 3 shows an example of a MIPS rediscovering the

```

1 def f(s):
2     a = 198;b = -11;c = -3;d = 483;e = 0;
3     ys = []
4     for i in range(20):
5         x = s[i]
6         next_a = -b+c+190
7         next_b = b-c-d-e+x+480
8         next_c = b-e+8
9         next_d = -b+e-x+472
10        next_e = a+b-e-187
11        a = next_a;b = next_b;c = next_c;d
12        = next_d;e = next_e;
13        y = -d+483
14        ys.append(y)
15    return ys
    
```

```

1 def f(s):
2     a = 0;b = 0;c = 0;d = 0;e = 0;
3     ys = []
4     for i in range(20):
5         x = s[i]
6         next_a = +x
7         next_b = a
8         next_c = b
9         next_d = c
10        next_e = d
11        a = next_a;b = next_b;c = next_c;d
12        = next_d;e = next_e;
13        y = a+b+c+d+e
14        ys.append(y)
15    return ys
    
```

Figure 4. Comparison of code generated from an RNN trained on Sum_Last5, with (top) and without (top) normalizers. The whitening normalizer provided numerical stability to the Jordan normal form normalizer, which itself simplified the recurrent portion of the program. The Toeplitz and de-biasing normalizers jointly sparsified the occurrences of x in the program, and the number of terms required to compute y . The quantization normalizer enabled all variables to be represented as integers.

“ripple-carry adder” algorithm. The normalizers significantly simplified some of the resulting programs, as illustrated in Fig. 4, and sometimes made the difference between MIPS failing and succeeding. We found that applying a small $L1$ weight regularization sometimes facilitated integer autoencoding by axis-aligning the lattice.

5. Conclusions

We have presented MIPS, a novel method for program synthesis based on automated mechanistic interpretability of neural networks trained to perform the desired task, auto-distilling the learned algorithm into Python code. Its essence is to first train a recurrent neural network to learn a clever finite state machine that performs the task, and then automatically figure out how this machine works.

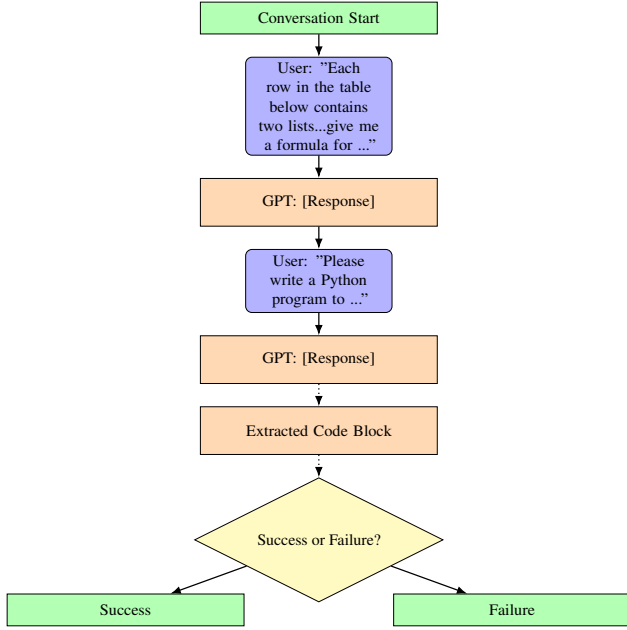


Figure 5. We compare MIPS against program synthesis with the large language model GPT-4 Turbo, prompted with a “chain-of-thought” approach. It begins with the user providing a task, followed by the model’s response, and culminates in assessing the success or failure of the generated Python code based on its accuracy in processing the provided lists.

5.1. Findings

We found MIPS highly complementary to LLM-based program synthesis with GPT-4 Turbo, with each approach solving many tasks that stumped the other. Whereas LLM-based methods have the advantage of drawing upon a vast corpus of human training data, MIPS has the advantage of discovering algorithms from scratch without human hints, with the potential to discover entirely new algorithms. As opposed to genetic programming approaches, MIPS leverages the power of deep learning by exploiting gradient information.

Program synthesis aside, our results shed further light on mechanistic interpretability, specifically on how neural networks represent bits and integers. We found that n integers tend to get encoded linearly in n dimensions, but generically in non-orthogonal directions with an additive offset. This is presumably because there are many more such messy encodings than simple ones, and the messiness can be easily (linearly) decoded. We saw that n bits sometimes get encoded as an n -dimensional parallelogram, but not always — possibly because linear decodability is less helpful when the subsequent bit operations to be performed are nonlinear anyway.

5.2. Outlook

Our work is merely a modest first attempt at mechanistic-interpretability-based program synthesis, and there are many obvious generalizations worth trying in future work. For example:

1. Improvements in training and integer autoencoding (since many of our failed examples failed only just barely)
2. Generalization from RNNs to other architectures such as transformers
3. Generalization from bits and integers to more general extractable data types such as floating-point numbers and various discrete mathematical structures and knowledge representations
4. Scaling to tasks requiring much larger neural networks
5. Automated formal verification of synthesized programs (we perform such verification with Dafny in Appendix F.1 to show that our MIPS-learned ripple adder correctly adds *any* binary numbers, not merely those in the test set, but such manual work should ideally be fully automated)

LLM-based coding co-pilots are already highly useful for program synthesis tasks based on verbal problem descriptions or auto-complete, and will only get better. MIPS instead tackles program synthesis based on test cases alone. This makes it analogous to *symbolic regression* (Udrescu et al., 2020; Cranmer, 2023), which has already proven useful for various science and engineering applications (Cranmer et al., 2020; Ma et al., 2022) where one wishes to approximate data relationships with symbolic formulae. The MIPS framework generalizes symbolic regression from feed-forward formulae to programs with loops, which are in principle Turing complete. If this approach can be scaled up, it may enable promising opportunities for making machine-learned algorithms more interpretable, verifiable, and trustworthy.

Broader Impact

Because machine-learned algorithms now outperform traditional human-discovered algorithms on many tasks, there are incentives to deploy them even without a full understanding of how they work and of whether they are biased, unsafe, or otherwise problematic. The aspirational broader impact motivating this paper is to help automate the process of making AI systems more transparent, robust, and trustworthy, with the ultimate goal of developing provably safe AI systems (Tegmark & Omohundro, 2023).

Acknowledgements

We thank Wes Gurnee, James Liu, and Armaun Sanaye for helpful conversations and suggestions. This work is supported by Erik Otto, Jaan Tallinn, the Rothberg Family Fund for Cognitive Science, the NSF Graduate Research Fellowship (Grant No. 2141064), and IAIFI through NSF grant PHY-2019786.

References

- Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., and Saunders, W. Language models can explain neurons in language models. <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>, 2023.
- Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.
- Cammarata, N., Goh, G., Carter, S., Schubert, L., Petrov, M., and Olah, C. Curve detectors. *Distill*, 2020. doi: 10.23915/distill.00024.003. <https://distill.pub/2020/circuits/curve-detectors>.
- Charton, F. Can transformers learn the greatest common divisor? *arXiv preprint arXiv:2308.15594*, 2023.
- Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*, 2023.
- Cranmer, M. Interpretable machine learning for science with pysr and symbolicregression. *jl. arXiv preprint arXiv:2305.01582*, 2023.
- Cranmer, M., Sanchez Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S. Discovering symbolic models from deep learning with inductive biases. *Advances in Neural Information Processing Systems*, 33: 17429–17442, 2020.
- Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.
- Goh, G., †, N. C., †, C. V., Carter, S., Petrov, M., Schubert, L., Radford, A., and Olah, C. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030. <https://distill.pub/2021/multimodal-neurons>.
- Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
- Gurnee, W. and Tegmark, M. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.
- Hanna, M., Liu, O., and Variengien, A. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *arXiv preprint arXiv:2305.00586*, 2023.
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.
- Lindner, D., Kramár, J., Farquhar, S., Rahtz, M., McGrath, T., and Mikulik, V. Tracr: Compiled transformers as a laboratory for interpretability. *arXiv preprint arXiv:2301.05062*, 2023.
- Liu, Z., Kitouni, O., Nolte, N., Michaud, E. J., Tegmark, M., and Williams, M. Towards understanding grokking: An effective theory of representation learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=6at6rB3IZm>.
- Ma, H., Narayanaswamy, A., Riley, P., and Li, L. Evolving symbolic density functionals. *Science Advances*, 8(36): eabq0279, 2022.
- Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.
- McGrath, T., Kaphishnikov, A., Tomašev, N., Pearce, A., Wattenberg, M., Hassabis, D., Kim, B., Paquet, U., and Kramnik, V. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119 (47):e2206625119, 2022.
- Nanda, N., Chan, L., Liberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

- Odena, A., Shi, K., Bieber, D., Singh, R., Sutton, C., and Dai, H. Bustle: Bottom-up program synthesis through learning-guided exploration. *arXiv preprint arXiv:2007.14381*, 2020.
- Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.
- Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.
- Quirke, P. et al. Understanding addition in transformers. *arXiv preprint arXiv:2310.13121*, 2023.
- Sobania, D., Briesch, M., and Rothlauf, F. Choose your programming copilot: A comparison of the program synthesis performance of github copilot and genetic programming. In *Proceedings of the genetic and evolutionary computation conference*, pp. 1019–1027, 2022.
- Syed, A., Rager, C., and Conmy, A. Attribution patching outperforms automated circuit discovery. *arXiv preprint arXiv:2310.10348*, 2023.
- Tegmark, M. and Omohundro, S. Provably safe systems: the only path to controllable agi. *arXiv preprint arXiv:2309.01933*, 2023.
- Toshniwal, S., Wiseman, S., Livescu, K., and Gimpel, K. Chess as a testbed for language model state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11385–11393, 2022.
- Udrescu, S.-M., Tan, A., Feng, J., Neto, O., Wu, T., and Tegmark, M. Ai feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. *Advances in Neural Information Processing Systems*, 33:4860–4871, 2020.
- Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.
- Wu, J., Wei, L., Jiang, Y., Cheung, S.-C., Ren, L., and Xu, C. Programming by example made easy. *ACM Transactions on Software Engineering and Methodology*, 33(1):1–36, 2023.
- Zhong, Z., Liu, Z., Tegmark, M., and Andreas, J. The clock and the pizza: Two stories in mechanistic explanation of neural networks. *arXiv preprint arXiv:2306.17844*, 2023.
- Zhou, B. and Ding, G. Survey of intelligent program synthesis techniques. In *International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2023)*, volume 12941, pp. 1122–1136. SPIE, 2023.

A. Lattice finding using generalized greatest common divisor (GCD)

Our method often encounters cases where hidden states secretly form an affine transformation of an integer lattice. However, not all lattice points are observed in training samples, so our goal is to recover the hidden integer lattice from sparse observations.

A.1. Problem formulation

Suppose we have a set of lattice points in \mathbb{R}^D spanned by D independent basis vectors, \mathbf{b}_i ($i = 1, 2, \dots, D$). Each lattice point j has the position

$$\mathbf{x}_j = \sum_{i=1}^D a_{ji} \mathbf{b}_i + \mathbf{c}, \quad (5)$$

where \mathbf{c} is a global translation vector, and the coefficients a_{ji} are integers

Our problem: given N such data points $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$, how can we recover the integer coefficients a_{ji} for each point data point as well as \mathbf{b}_i and \mathbf{c} ?

Note that even when the whole lattice is given, there are still degrees of freedom for the solution. For example, $\{\mathbf{c} \mapsto \mathbf{c} + \mathbf{b}_i, a_{ji} \mapsto a_{ji} - 1\}$ remains a solution, and $\{\mathbf{b}_i \rightarrow \sum_{j=1}^D \Lambda_{ij} \mathbf{b}_j\}$ remains a solution if Λ is an integer matrix whose determinant is ± 1 . So our success criterion is: (1) a_{ji} are integers; (2) the discovered bases and the true bases have the same determinant (the volume of a unit cell). Once a set of bases is found, we can simplify them by minimizing their total norms over valid transformations ($\Lambda \in \mathbb{Z}^{D \times D}, \det(\Lambda) = \pm 1$).

A.2. Regular GCD

As a reminder, given a list of n numbers $\{y_1, y_2, \dots, y_n\}$, a common divisor d is a number such that for all i , $\frac{y_i}{d}$ is an integer. All common divisors are the set $\{d | y_i/d \in \mathbb{Z}, \text{ and the greatest common divisor (GCD) is the largest number in this set. Because}$

$$\text{GCD}(y_1, \dots, y_n) = \text{GCD}(y_1, \text{GCD}(y_2, \text{GCD}(y_3, \dots))), \quad (6)$$

it without loss of generality suffices to consider the case $n = 2$. A common algorithm to compute GCD of two number is the so-called Euclidean algorithm. We start with two numbers r_0, r_1 and $r_0 > r_1$, which is step 0. For the k^{th} step, we perform division-with-remainder to find the quotient q_k and the remainder r_k so that $r_{k-2} = q_k r_{k-1} + r_k$ with $|r_{k-1}| > |r_k|$ ¹. The algorithm will eventually produce a zero remainder $r_N = 0$, and the other non-zero remainder r_{N-1} is the greatest common divisor. For example, $\text{GCD}(55, 45) = 5$, because

$$\begin{aligned} 55 &= 1 \times 45 + 10, \\ 45 &= 4 \times 10 + 5, \\ 10 &= 2 \times 5 + 0. \end{aligned} \quad (7)$$

A.3. Generalized GCD in D dimensions

Given a list of n vectors $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$ where $\mathbf{y}_i \in \mathbb{R}^D$, and assuming that these vectors are in the lattice described by Eq. (5), we can without loss of generality set $\mathbf{c} = 0$, since we can always redefine the origin. In D dimensions, the primitive of interest is the D -dimensional parallelogram: a line segment for $D = 1$ (one basis vector), a parallelogram for $D = 2$ (two basis vectors), parallelepiped for $D = 3$ (three basis vectors), *etc.*

One can construct a D -dimensional parallelogram by constructing its basis vectors as a linear integer combination of \mathbf{y}_j , i.e.,

$$\mathbf{q}_i = \sum_{j=1}^n m_{ij} \mathbf{y}_j, m_{ij} \in \mathbb{Z}, i = 1, 2, \dots, D. \quad (8)$$

The goal of D -dimensional GCD is to find a “minimal” parallelogram, such that its volume (which is $\det(\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_D)$)

¹We are considering a general case where r_0 and r_1 may be negative. Otherwise r_k can always be positive numbers, hence no need to use the absolute function.

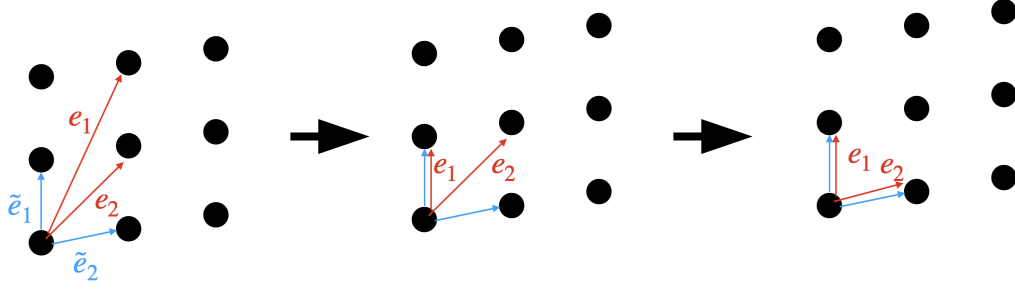


Figure 6. Both red and blue basis form a minimal parallelogram (in terms of cell volume), but one can further simplify red to blue by linear combination (simplicity in the sense of small ℓ_2 norm).

is GCD of volumes of other possible parallelograms. Once the minimal parallelogram is found², we can also determine \mathbf{b}_i in Eq. (5), since \mathbf{b}_i are exactly \mathbf{q}_i ! To find the minimal parallelogram, we need two steps: (1) figure out the unit volume; (2) figure out $\mathbf{q}_i (i = 1, 2, \dots)$ whose volume is the unit volume.

Step 1: Compute unit volume V_0 . We first define *representative* parallelograms as one where all $i = 1, 2, \dots, D$, $\mathbf{m}_i \equiv (m_{i1}, m_{i2}, \dots, m_{iD})$ are one-hot vectors, *i.e.*, with only one element being 1 and 0 otherwise. It is easy to show that the volume of any parallelogram is a linear integer combination of volumes of representative parallelograms, so WLOG we can focus on representative parallelograms. We compute the volumes of all representative parallelograms, which gives a volume array. Since volumes are just scalars, we can get the unit volume V_0 by calling the regular GCD of the volume array.

Step 2: Find a minimal parallelogram (whose volume is the unit volume computed in step 1). Recall that in regular GCD, we are dealing with two numbers (scalars). To leverage this in the vector case, we need to create scalars out of vectors, and make sure that the vectors share the same linear structure as the scalars so that we can extend division-and-remainder to vectors. A natural scalar is volume. Now consider two parallelograms P1 and P2, which share $D - 1$ basis vectors $(\mathbf{y}_3, \dots, \mathbf{y}_{D+1})$, but last basis vector is different: \mathbf{y}_1 for P1 and \mathbf{y}_2 for P2. Denote their volume as V_1 and V_2 :

$$\begin{aligned} V_1 &= \det(\mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_4, \dots, \mathbf{y}_D) \\ V_2 &= \det(\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \dots, \mathbf{y}_D) \end{aligned} \quad (9)$$

Since

$$aV_1 + bV_2 = \det(a\mathbf{y}_1 + b\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \dots, \mathbf{y}_D), \quad (10)$$

which shows that (V_1, V_2) and $(\mathbf{y}_1, \mathbf{y}_2)$ share the same linear structure. We can simply apply division-and-remainder to V_1 and V_2 as in regular GCD:

$$V'_1, V'_2 = \text{GCD}(V_1, V_2), \quad (11)$$

whose quotients in all iterations are saved and transferred to \mathbf{y}_1 and \mathbf{y}_2 :

$$\mathbf{y}'_1, \mathbf{y}'_2 = \text{GCD_with_predefined_quotients}(\mathbf{y}_1, \mathbf{y}_2). \quad (12)$$

If $V_1 = V_0$ (which is the condition for minimal parallelogram), the algorithm terminates and returns $(\mathbf{y}'_1, \mathbf{y}_3, \mathbf{y}_4, \dots, \mathbf{y}_D)$. If $V_1 > V_0$, we need to repeat step 2 with the new vector list $\{\mathbf{y}'_1, \mathbf{y}_3, \dots, \mathbf{y}_N\}$.

Why can we remove \mathbf{y}'_2 for next iteration? Note that although eventually $V'_1 > 0$ and $V'_2 = 0$, typically $\mathbf{y}_2 \neq 0$. However, since

$$0 = V'_2 = \det(\mathbf{y}'_2, \mathbf{y}_3, \mathbf{y}_4, \dots, \mathbf{y}_D), \quad (13)$$

this means \mathbf{y}'_2 is a linear combination of $(\mathbf{y}_3, \dots, \mathbf{y}_D)$, hence can be removed from the vector list.

Step 3: Simplification of basis vectors. We want to further simplify basis vectors. For example, the basis vectors obtained in step 2 may have large norms. For example $D = 2$, the standard integer lattice has $\mathbf{b}_1 = (1, 0)$ and $\mathbf{b}_2 = (0, 1)$, but they

²There could be many minimal parallelograms, but finding one is sufficient.

are infinitely many possibilities after step 2, as long as $pt - sq = \pm 1$ for $\mathbf{b}_1 = (p, q)$ and $\mathbf{b}_2 = (s, t)$, e.g., $\mathbf{b}_1 = (3, 5)$ and $\mathbf{b}_2 = (4, 7)$.

To minimize ℓ_2 norms, we choose a basis and project-and-subtract for other bases. Note that: (1) again we are only allowed to subtract integer times of the chosen basis; (2) the volume of the parallelogram does not change since the project-and-subtract matrix has determinant 1 (suppose $\mathbf{b}_i (i = 2, 3, \dots, D)$ are projected to \mathbf{b}_1 and subtracted by multiples of \mathbf{b}_1 . p_* represents projection integers):

$$\begin{pmatrix} 1 & p_{2 \rightarrow 1} & p_{3 \rightarrow 1} & \cdots & p_{D \rightarrow 1} \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \quad (14)$$

We do this iteratively, until no norm can become shorter via project-and-subtract. Please see Figure 6 for an illustration of how simplification works for a 2D example.

Computation overhead is actually surprisingly small. In typical cases, we only need to call $O(1)$ times of GCD.

Dealing with noise Usually the integer lattice in the hidden space is not perfect, i.e., vulnerable to noise. How do we extract integer lattices in a robust way in the presence of noise? Note that the terminating condition for the GCD algorithm is when the remainder is exactly zero - we relax this condition to that the absolute value of the remainder to be smaller than a threshold ϵ_{gcd} . Another issue regarding noise is that noise can be accumulated in the GCD iterations, so we hope that GCD can converge in a few steps. To achieve this, we select hidden states in a small region with data fraction $p\%$ of the whole data. Since both ϵ_{gcd} and p depends on data and neural network training which we do not know a priori, we choose to grid sweep $\epsilon_{\text{gcd}} \in [10^{-3}, 1]$ and $p \in (0.1, 100)$; for each $(\epsilon_{\text{gcd}}, p)$, we obtain an integer lattice and compute its description length. We select the $(\epsilon_{\text{gcd}}, p)$ which gives the lattice with the smallest description length. The description length includes two parts: integer descriptions of hidden states $\log(1 + |Z|^2)$, and residual of reconstruction $\log(1 + (\frac{AZ+b-X}{\epsilon_{\text{dl}}})^2)$ with $\epsilon_{\text{dl}} = 10^{-4}$.

B. Linear lattice finder

Although our RNN can represent general nonlinear functions, in the special case when the RNN actually performs linear functions, program synthesis can be much easier. So if the hidden MLP is linear, we would expect the hidden states to be an integer lattice, because inputs are integer lattices and the mappings are linear. Effectively the hidden MLP works as a linear function: $\mathbf{h}^{(t)} = W_h \mathbf{h}^{(t-1)} + W_i \mathbf{x}^{(t)}$ (we neglected the bias term since it is not relevant to finding basis vectors of a lattice).

Suppose we have input series $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(t)}$, then $\mathbf{h}^{(t)}$ is

$$\mathbf{h}^{(t)} = \sum_{j=1}^t W_h^{t-j} W_i \mathbf{x}_j, \quad (15)$$

Since \mathbf{x}_j themselves are integer lattices, we could then interpret the following as basis vectors:

$$W_h^{t-j} W_i, j = 1, 2, \dots, t, \quad (16)$$

which are not necessarily independent. For example, for the task of summing up the last two numbers, $W_h W_i$ and W_i are non-zero vectors and are independent, while others $W_h^n W_i \approx 0, n \geq 2$. Then $W_h W_i$ and W_i are the two basis vectors for the lattice. In general, we measure the norm of all the candidate basis vectors, and select the first k vectors with highest norms, which are exactly basis vectors of the hidden lattice.

C. Symbolic regression

The formulation of symbolic regression is that one has data pair $(\mathbf{x}_i, y_i), i = 1, 2, \dots, N$ with N data samples. The goal is to find a symbolic formula f such that $y_i = f(\mathbf{x}_i)$. A function is expressed in reverse polish notation (RPN), for example, $|a| - c$ is expressed as $\text{aAc}-$ where A stands for the absolute value function. We have three types of variables:

- type-0 operator. We include input variables and constants.
- type-1 operator (takes in one type-0 to produce one type-0). We include operations $\{>, <, \sim, H, D, A\}$. $>$ means $+1$; $<$ means -1 ; \sim means negating the number; D is dirac delta which outputs 1 only when taking in 0; A is the absolute value function;
- type-2 operator (takes in two type-0 to produce one type-0). We include operations $\{+, *, -, \%\}$. $+$ means addition of two numbers; $*$ means multiplication of two numbers; $-$ means subtraction of two numbers; $\%$ is the remainder of one number module the other.

There are only certain types of templates (a string of numbers consisting of 0,1,2) that are syntactically correct. For example, 002 is correct while 02 is incorrect. We iterate over all the templates not longer than 6 symbols, and for each template, we try all the variable combinations. Each variable combination corresponds to a symbolic equation f , for which we can check whether $f(\mathbf{x}_i) = y_i$ for 100 data points. If success, we terminate the brute force program and return the successful formula. If brute force search does not find any correct symbolic formula within compute budget, we will simply return the formula a , to make sure that the program can still be synthesized but simply fail to make correct predictions.

D. Neural Network Normalization Algorithms

It is well known that neural networks exhibit a large amount of symmetry. That is, there are many transformations that can be applied to networks without affecting the map $y = f(x)$ that they compute. A classic example is to permute the neurons within layers.

In this section, we describe a suite of normalizers that we use to transform our networks into a standard form, such that the algorithms that they learn are easier to interpret. We call our five normalizers “Whitening”, “Jordan normal form (JNF)”, “Toeplitz”, “De-bias”, and “Quantization”.

The main symmetry which we focus on is a linear transformation of the hidden space $\mathbf{h} \mapsto \mathbf{A}\mathbf{h}$, which requires the following changes to f and g :

$$\begin{aligned} f(\mathbf{h}, \mathbf{x}) = \mathbf{W}\mathbf{h} + \mathbf{V}\mathbf{x} + \mathbf{b} &\implies f(\mathbf{h}, \mathbf{x}) = \mathbf{A}\mathbf{W}\mathbf{A}^{-1}\mathbf{h} + \mathbf{A}\mathbf{V}\mathbf{x} + \mathbf{A}\mathbf{b} \\ g(\mathbf{h}) = G(\mathbf{U}\mathbf{h} + \mathbf{c}) &\implies f(\mathbf{h}) = G(\mathbf{U}\mathbf{A}^{-1}\mathbf{h} + \mathbf{c}) \end{aligned}$$

and is implemented by changing the weights:

$$\begin{aligned} \mathbf{W} &\implies \mathbf{A}\mathbf{W}\mathbf{A}^{-1} \\ \mathbf{V} &\implies \mathbf{A}\mathbf{V} \\ \mathbf{b} &\implies \mathbf{A}\mathbf{b} \\ \mathbf{U} &\implies \mathbf{U}\mathbf{A}^{-1} \end{aligned}$$

For this symmetry, we can apply an arbitrary invertible similarity transformation \mathbf{A} to \mathbf{W} , which is the core idea underlying our normalizers, three of which have their own unique ways of constructing \mathbf{A} , as we describe in the sections below. Most importantly, one of our normalizers exploits \mathbf{A} to convert the hidden-to-hidden transformation \mathbf{W} into Jordan normal form, in the case where f is linear. Recent work has shown that large recurrent networks with linear hidden-to-hidden transformations, such as state space models (Gu & Dao, 2023) can perform just as well as transformer-based models in language modeling on a large scale. A main advantage of using linear hidden-to-hidden transformations is the possibility of expressing the hidden space in it’s eigenbasis. This causes the hidden-to-hidden transformation to become diagonal, so that it can be computed more efficiently. In practice, modern state space models assume diagonality, and go further to assume the elements on the diagonal are real; they fix the architecture to be this way during training.

By doing this, we ignore the possibility of linear hidden-to-hidden transformations that cannot be transformed into a real diagonal matrix via diagonalization. Such examples include rotation matrices (whose eigenvalues may be complex), and shift matrices (whose eigenvalues are degenerate and whose eigenvectors are duplicated). A more general form than the diagonal form is the Jordan normal form, which consists of Jordan blocks along the diagonal, each of which has the form $\lambda\mathbf{I} + \mathbf{S}$ for an eigenvalue λ and the shift matrix \mathbf{S} with ones on the superdiagonal and zeros elsewhere. The diagonalization

is a special case of Jordan normal form, and all matrices can be transformed to Jordan normal form. A simple transformation can also be applied to Jordan normal forms that contain pairs of complex generalized eigenvectors, to convert them into real matrices.

For nonlinear hidden-to-hidden transformations, we compute \mathbf{W} as though the nonlinearities have been removed.

D.1. Whitening Transformation

Similar to normalizing the means and variances of a dataset before feeding it into a machine learning model, a good first preprocessing step is to normalize the distribution of hidden states. We therefore choose to apply a whitening transformation to the hidden space. To compute the transformation, we compute the covariance matrix of hidden activations across the dataset, and use the singular value decomposition (SVD) of this covariance matrix to find the closest transformation to the identity that will bring this covariance matrix to the identity. We ignore any directions with covariance less than $\epsilon = 0.1$, which cause more instability when normalized. We then post-apply this transformation to the last linear layer of the hidden-to-hidden transformation and its biases, and pre-apply its inverse to the first layers of the hidden-to-hidden and hidden-to-output transformations. This leaves the net behavior of the network unchanged. Other transformations which we use in other normalizers operate in a similar manner, by post-applying and pre-applying a transformation and its inverse transformation to the first and last layers that interact with the hidden space.

D.2. Jordan Normal Form Transformation

Critically, the hidden-to-hidden transformations which we would like to convert into Jordan normal form are imperfect because they are learned. Eigenvectors belonging to each Jordan block must be identical, whereas this will only be approximately true of the learned transformation.

The Jordan normal form of a matrix is unstable; consider a matrix

$$\mathbf{W} = \begin{pmatrix} 0 & 1 \\ \delta & 0 \end{pmatrix}$$

which, when $\delta \neq 0$, can be transformed into Jordan normal form by:

$$\begin{pmatrix} 0 & 1 \\ \delta & 0 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 \\ \sqrt{\delta} & -\sqrt{\delta} \end{pmatrix}}_{\mathbf{T}} \begin{pmatrix} \sqrt{\delta} & 0 \\ 0 & -\sqrt{\delta} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ \sqrt{\delta} & -\sqrt{\delta} \end{pmatrix}^{-1} \quad (17)$$

but when $\delta = 0$, is transformed into Jordan normal form by:

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{\mathbf{T}} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1} \quad (18)$$

As we can see, all of the matrices in the decomposition are unstable near $\delta = 0$, so the issue of error thresholding is not only numerical, but is mathematical in nature as well.

We would like to construct an algorithm which computes the Jordan normal form with an error threshold $|\delta| < \epsilon = 0.7$ within which the algorithm will pick the transformation \mathbf{T} from Equation (18) instead of from Equation (17).

Our algorithm first computes the eigenvalues λ_i , and then iteratively solves for the generalized eigenvectors which lie in $\ker((\mathbf{W} - \lambda\mathbf{I})^k)$ for increasing k . The approximation occurs whenever we compute the kernel (of unknown dimension) of a matrix \mathbf{X} ; we take the SVD of \mathbf{X} and treat any singular vectors as part of the nullspace if their singular values are lower than the threshold ϵ , calling the result $\epsilon\text{-ker}(\mathbf{X})$.

Spaces are always stored in the form of a rectangular matrix \mathbf{F} of orthonormal vectors, and their dimension is always the width of the matrix. We build projections using $\text{proj}(\mathbf{F}) = \mathbf{F}\mathbf{F}^H$, where \mathbf{F}^H denotes the conjugate transpose of \mathbf{F} . We compute kernels $\ker(\mathbf{X})$ of known dimension of matrices \mathbf{X} by taking the SVD $\mathbf{X} = \mathbf{V}_1\mathbf{S}\mathbf{V}_2^H$ and taking the last singular vectors in \mathbf{V}_2^H . We compute column spaces of projectors of known dimension by taking the top singular vectors of the SVD.

The steps in our algorithm are as follows:

1. Solve for the eigenvalues λ_i of \mathbf{W} , and check that eigenvalues that are within ϵ of each other form groups, ie. that $|\lambda_i - \lambda_j| \leq \epsilon$ and $|\lambda_j - \lambda_k| \leq \epsilon$ always implies $|\lambda_k - \lambda_i| \leq \epsilon$. Compute the mean eigenvalue for every group.
2. Solve for the approximate kernels of $\mathbf{W} - \lambda\mathbf{I}$ for each mean eigenvalue λ . We will denote this operation by $\epsilon\text{-ker}(\mathbf{W} - \lambda\mathbf{I})$. We represent these kernels by storing the singular vectors whose singular values are lower than ϵ . Also, construct a “corrected matrix” of $\mathbf{W} - \lambda\mathbf{I}$ for every λ by taking the SVD, discarding the low singular values, and multiplying the pruned decomposition back together again.
3. Solve for successive spaces \mathbf{F}_k of generalized eigenvectors at increasing depths k along the set of Jordan chains with eigenvalue λ , for all λ . In other words, find chains of mutually orthogonal vectors which are mapped to zero after exactly k applications of the map $\mathbf{W} - \lambda\mathbf{I}$. We first solve for $\mathbf{F}_0 = \ker(\mathbf{W} - \lambda\mathbf{I})$. Then for $k > 0$, we first solve for $\mathbf{J}_k = \epsilon\text{-ker}((\mathbf{I} - \text{proj}(\mathbf{F}_{k-1}))(\mathbf{W} - \lambda\mathbf{I}))$ and deduce the number of chains which reach depth k from the dimension of \mathbf{J}_k , and then solve for $\mathbf{F}_k = \text{col}(\text{proj}(\mathbf{J}_k) - \text{proj}(\mathbf{F}_0))$.
4. Perform a consistency check to verify that the dimensions of \mathbf{F}_k always stay the same or decrease with k . Go through the spaces \mathbf{F}_k in reverse order, and whenever the dimension of \mathbf{F}_k decreases, figure out which direction(s) are not mapped to by applying $\mathbf{W} - \lambda\mathbf{I}$ to \mathbf{F}_{k+1} . Do this by building a projector \mathbf{J} from mapping vectors representing \mathbf{F}_{k+1} through $\mathbf{W} - \lambda\mathbf{I}$, and taking $\text{col}(\text{proj}(\mathbf{F}_k) - \mathbf{J})$. Solve for the Jordan chain by repeatedly applying $\text{proj}(\mathbf{F}_i)(\mathbf{W}_i - \lambda\mathbf{I})$ for i starting from $k - 1$ and going all the way down to zero.
5. Concatenate all the Jordan chains together to form the transformation matrix \mathbf{T} .

The transformation \mathbf{T} consists of generalized eigenvectors which need not be completely real but may also include pairs of generalized eigenvectors that are complex conjugates of each other. Since we do not want the weights of our normalized network to be complex, we also apply a unitary transformation which changes any pair of complex generalized eigenvectors into a pair of real vectors, and the resulting block of \mathbf{W} into a multiple of a rotation matrix. As an example, for a real 2 by 2 matrix \mathbf{W} with complex eigenvectors, we have

$$\begin{aligned}\mathbf{W} &= \mathbf{T} \begin{pmatrix} a + bi & 0 \\ 0 & a - bi \end{pmatrix} \mathbf{T}^{-1} \\ &= \mathbf{T}\mathbf{T}' \begin{pmatrix} a & -b \\ b & a \end{pmatrix} (\mathbf{T}\mathbf{T}')^{-1}, \quad \mathbf{T}' = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix}\end{aligned}$$

D.3. Toeplitz Transformation

Once \mathbf{W} is in Jordan normal form, each Jordan block is an upper triangular Toeplitz matrix. Upper-triangular Toeplitz matrices, including Jordan blocks, will always commute with each other, because they are all polynomials of the shift matrix (which has ones on the superdiagonal and zeros elsewhere,) and therefore these transformations will leave \mathbf{W} unchanged, but will still affect \mathbf{V} . We split \mathbf{V} up into parts operated on by each Jordan block, and use these Toeplitz transformations to reduce the most numerically stable columns of each block of \mathbf{V} to one-hot vectors. The numerical stability of a column vector is determined by the absolute value of the bottom element of that column vector, since it’s inverse will become the degenerate eigenvalues of the resulting Toeplitz matrix. If no column has a numerical stability above $\epsilon = 0.0001$, we pick the identity matrix for our Toeplitz transformation.

D.4. De-biasing Transformation

Oftentimes, \mathbf{W} is not full rank, and has a nontrivial nullspace. The bias \mathbf{b} will have some component in the direction of this nullspace, and eliminating this component only affects the behavior of the output network g , and the perturbation cannot carry on to the remainder of the sequence via f . Therefore, we eliminate any such component, and compensate accordingly by modifying the bias in the first affine layer of g . We identify the nullspaces by taking an SVD and identifying components whose singular value is less than $\epsilon = 0.1$.

D.5. Quantization Transformation

After applying all of the previous transformations to the RNN, it is common for many of the weights to become close to zero or some other small integer. Treating this as a sign that the network is attempting to implement discrete operations using integers, we snap any weights and biases that are within a threshold $\epsilon = 0.01$ of an integer, to that integer. For certain simple tasks, sometimes this allows the entire network to become quantized.

E. Supplementary training data details

Here we present additional details on the benchmark tasks marked "see text" in Table 1:

- **Div_3/5/7:** This is a long division task for binary numbers. The input is a binary number, and the output is that binary number divided by 3, 5, or 7, respectively. The remainder is discarded. For example, we have $1000011/11=0010110$ ($67/3=22$). The most significant bits occur first in the sequence.
- **Dithering:** This is a basic image color quantization task, for 1D images. We map 4-bit images to 1-bit images such that the cumulative sum of pixel brightnesses of both the original and dithered images remains as close as possible.
- **Newton_Gravity:** This is an euler forward propagation technique which follows the equation $F = input - 1, v \mapsto v + F, x \mapsto x + v$.
- **Newton_Spring:** This is an euler forward propagation technique which follows the equation $F = input - x, v \mapsto v + F, x \mapsto x + v$.
- **Newton_Magnetic:** This is an euler forward propagation technique which follows the equation $F_x = input_1 - v_y, F_y = input_2 + v_x, \mathbf{v} \mapsto \mathbf{v} + \mathbf{F}, \mathbf{x} \mapsto \mathbf{x} + \mathbf{v}$.

F. Generated programs

This section includes all successfully generated Python programs.

Binary-Addition

```

1
2 def f(s,t):
3     a = 0;b = 0;
4     ys = []
5     for i in range(10):
6         c = s[i]; d = t[i];
7         next_a = b ^ c ^ d
8         next_b = b+c+d>1
9         a = next_a;b = next_b;
10        y = a
11        ys.append(y)
12    return ys
    
```

Bitwise-Or

```

1
2 def f(s,t):
3     a = 0;
4     ys = []
5     for i in range(10):
6         b = s[i]; c = t[i];
7         next_a = b+c>0
8         a = next_a;
9         y = a
10        ys.append(y)
11    return ys
    
```

Bitwise-Xor

```

1
2 def f(s,t):
3     a = 0;
4     ys = []
5     for i in range(10):
6         b = s[i]; c = t[i];
7         next_a = b ^ c
8         a = next_a;
9         y = a
10        ys.append(y)
11    return ys
    
```

Bitwise-And

```

1
2 def f(s,t):
3     a = 0;b = 1;
4     ys = []
5     for i in range(10):
6         c = s[i]; d = t[i];
7         next_a = (not a and not b and c and
8         d) or (not a and b and not c and d) or
9         (not a and b and c and not d) or (not
10        a and b and c and d) or (a and not b
11        and c and d) or (a and b and c and d)
12        next_b = c+d==0 or c+d==2
13        a = next_a;b = next_b;
14        y = a+b>1
15        ys.append(y)
16    return ys
    
```

Bitwise-Not

```

1
2 def f(s):
3     a = 1;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = x
8         a = next_a;
9         y = -a+1
10        ys.append(y)
11    return ys

```

Parity-Zeros

```

1
2 def f(s):
3     a = 0;
4     ys = []
5     for i in range(10):
6         b = s[i]
7         next_a = a+b==0 or a+b==2
8         a = next_a;
9         y = a
10        ys.append(y)
11    return ys

```

Parity-Last2

```

1
2 def f(s):
3     a = 0;b = 0;
4     ys = []
5     for i in range(10):
6         c = s[i]
7         next_a = c
8         next_b = a ^ c
9         a = next_a;b = next_b;
10        y = b
11        ys.append(y)
12    return ys

```

Sum-All

```

1
2 def f(s):
3     a = 884;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = a-x
8         a = next_a;
9         y = -a+884
10        ys.append(y)
11    return ys

```

Parity-Last3

```

1
2 def f(s):
3     a = 0;b = 0;c = 0;
4     ys = []
5     for i in range(10):
6         d = s[i]
7         next_a = d
8         next_b = c
9         next_c = a
10        a = next_a;b = next_b;c = next_c;
11        y = a ^ b ^ c
12        ys.append(y)
13    return ys

```

Sum-Last2

```

1
2 def f(s):
3     a = 0;b = 99;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = -b+x+99
8         next_b = -x+99
9         a = next_a;b = next_b;
10        y = a
11        ys.append(y)
12    return ys

```

Parity-All

```

1
2 def f(s):
3     a = 0;
4     ys = []
5     for i in range(10):
6         b = s[i]
7         next_a = a ^ b
8         a = next_a;
9         y = a
10        ys.append(y)
11    return ys

```

Sum-Last3

```

1
2 def f(s):
3     a = 0;b = 198;c = 0;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = x
8         next_b = -a-x+198
9         next_c = -b+198
10        a = next_a;b = next_b;c = next_c;
11        y = a+c
12        ys.append(y)
13    return ys

```


Sum-Last4

```

1
2 def f(s):
3     a = 0;b = 99;c = 0;d = 99;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = c
8         next_b = -x+99
9         next_c = -b-d+198
10        next_d = b
11        a = next_a;b = next_b;c = next_c;d
12        = next_d;
13        y = a-b-d+198
14        ys.append(y)
15    return ys

```

Sum-Last5

```

1
2 def f(s):
3     a = 198;b = -10;c = -2;d = 482;e = 1;
4     ys = []
5     for i in range(20):
6         x = s[i]
7         next_a = -b+c+190
8         next_b = b-c-d-e+x+480
9         next_c = b-e+8
10        next_d = -b+e-x+472
11        next_e = a+b-e-187
12        a = next_a;b = next_b;c = next_c;d
13        = next_d;e = next_e;
14        y = -d+483
15        ys.append(y)
16    return ys

```

Sum-Last6

```

1
2 def f(s):
3     a = 0;b = 295;c = 99;d = 0;e = 297;f =
4     99;
5     ys = []
6     for i in range(20):
7         x = s[i]
8         next_a = -b+295
9         next_b = b-c+f
10        next_c = b-c+d-97
11        next_d = -f+99
12        next_e = -a+297
13        next_f = -x+99
14        a = next_a;b = next_b;c = next_c;d
15        = next_d;e = next_e;f = next_f;
16        y = -b+c-e-f+592
17        ys.append(y)
18    return ys

```

Sum-Last7

```

1
2 def f(s):
3     a = 297;b = 198;c = 0;d = 99;e = 0;f =
4     -15;g = 0;
5     ys = []
6     for i in range(20):
7         x = s[i]
8         next_a = -a+d-f+g+480
9         next_b = a-d
10        next_c = d+e-99
11        next_d = -c+99
12        next_e = -b+198
13        next_f = -c+f+x
14        next_g = x
15        a = next_a;b = next_b;c = next_c;d
16        = next_d;e = next_e;f = next_f;g =
17        next_g;
18        y = -d+f+114
19        ys.append(y)
20    return ys

```

Current-Number

```

1
2 def f(s):
3     a = 99;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = -x+99
8         a = next_a;
9         y = -a+99
10        ys.append(y)
11    return ys

```

Prev1

```

1
2 def f(s):
3     a = 0;b = 99;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = -b+99
8         next_b = -x+99
9         a = next_a;b = next_b;
10        y = a
11        ys.append(y)
12    return ys

```

Prev2

```

1
2 def f(s):
3     a = 99;b = 0;c = 0;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = -x+99
8         next_b = -a+99
9         next_c = b
10        a = next_a;b = next_b;c = next_c;
11        y = c
12        ys.append(y)
13    return ys

```

Prev3

```

1
2 def f(s):
3     a = 0;b = 0;c = 99;d = 99;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = b
8         next_b = -c+99
9         next_c = d
10        next_d = -x+99
11        a = next_a;b = next_b;c = next_c;d
12        = next_d;
13        y = a
14        ys.append(y)
15    return ys

```

Prev4

```

1
2 def f(s):
3     a = 0;b = 99;c = 0;d = 99;e = 0;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = c
8         next_b = -a+99
9         next_c = -d+99
10        next_d = -e+99
11        next_e = x
12        a = next_a;b = next_b;c = next_c;d
13        = next_d;e = next_e;
14        y = -b+99
15        ys.append(y)
16    return ys

```

Prev5

```

1
2 def f(s):
3     a = 0;b = 0;c = 99;d = 99;e = 99;f =
4     99;
5     ys = []
6     for i in range(20):
7         x = s[i]
8         next_a = -c+99
9         next_b = -d+99
10        next_c = -b+99
11        next_d = e
12        next_e = f
13        next_f = -x+99
14        a = next_a;b = next_b;c = next_c;d
15        = next_d;e = next_e;f = next_f;
16        y = a
17        ys.append(y)
18    return ys

```

Previous-Equals-Current

```

1
2 def f(s):
3     a = 0;b = 0;
4     ys = []
5     for i in range(10):
6         c = s[i]
7         next_a = delta(c-b)
8         next_b = c
9         a = next_a;b = next_b;
10        y = a
11        ys.append(y)
12    return ys

```

Diff-Last2

```

1
2 def f(s):
3     a = 199;b = 100;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = -a-b+x+498
8         next_b = a+b-199
9         a = next_a;b = next_b;
10        y = a-199
11        ys.append(y)
12    return ys

```

Abs-Diff

```

1
2 def f(s):
3     a = 100;b = 100;
4     ys = []
5     for i in range(10):
6         c = s[i]
7         next_a = b
8         next_b = c+100
9         a = next_a;b = next_b;
10        y = abs(b-a)
11        ys.append(y)
12    return ys

```

Abs-Current

```

1
2 def f(s):
3     a = 0;
4     ys = []
5     for i in range(10):
6         b = s[i]
7         next_a = abs(b)
8         a = next_a;
9         y = a
10        ys.append(y)
11    return ys

```

Newton-Freebody

```

1
2 def f(s):
3     a = 82;b = 393;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = a-x
8         next_b = -a+b+82
9         a = next_a;b = next_b;
10        y = -a+b-311
11        ys.append(y)
12    return ys

```

Bit-Shift-Right

```

1
2 def f(s):
3     a = 0;b = 1;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = -b+1
8         next_b = -x+1
9         a = next_a;b = next_b;
10        y = a
11        ys.append(y)
12    return ys

```

Newton-Gravity

```

1
2 def f(s):
3     a = 72;b = 513;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = a-x+1
8         next_b = -a+b+x+71
9         a = next_a;b = next_b;
10        y = b-513
11        ys.append(y)
12    return ys

```

Bit-Dot-Prod-Mod2

```

1
2 def f(s,t):
3     a = 0;
4     ys = []
5     for i in range(10):
6         b = s[i]; c = t[i];
7         next_a = (not a and b and c) or (a
8         and not b and not c) or (a and not b
9         and c) or (a and b and not c)
10        a = next_a;
11        y = a
12        ys.append(y)
13    return ys

```

Newton-Spring

```

1
2 def f(s):
3     a = 64;b = 57;
4     ys = []
5     for i in range(10):
6         x = s[i]
7         next_a = a+b-x-57
8         next_b = -a+121
9         a = next_a;b = next_b;
10        y = -a+64
11        ys.append(y)
12    return ys

```

Add-Mod-3

```

1
2 def f(s):
3     a = 0;
4     ys = []
5     for i in range(10):
6         b = s[i]
7         next_a = (b+a)%3
8         a = next_a;
9         y = a
10        ys.append(y)
11    return ys

```

F.1. Formal Verification

The Dafny programming language is designed so that programs can be formally verified for correctness. The desired behavior of a program can be explicitly specified via preconditions, postconditions, and invariants, which are verified via automated theorem proving. These capabilities make Dafny useful in fields where correctness and safety are crucial.

We leverage Dafny’s robust verification capabilities to prove the correctness of the bit addition Python program synthesized by MIPS. The bit addition Python program was first converted to Dafny, then annotated with specific assertions, preconditions, and postconditions that defined the expected behavior of the code. Each annotation in the code was then formally verified by Dafny, ensuring that under all possible valid inputs, the code’s output would be consistent with the expected behavior. On line 79, we show that the algorithm found by MIPS is indeed equivalent to performing bit addition with length 10 bitvectors in Dafny.

Dafny-Code

```

1
2 function ArrayToBv10(arr: array<bool>): bv10 // Converts boolean array to bitvector
3   reads arr
4   requires arr.Length == 10
5   {
6     ArrayToBv10Helper(arr, arr.Length - 1)
7   }
8
9 function ArrayToBv10Helper(arr: array<bool>, index: nat): bv10
10  reads arr
11  requires arr.Length == 10
12  requires 0 <= index < arr.Length
13  decreases index
14  ensures forall i :: 0 <= i < index ==> ((ArrayToBv10Helper(arr, i) >> i) & 1) == (if arr
15    [i] then 1 else 0)
16  {
17    if index == 0 then
18      (if arr[0] then 1 else 0) as bv10
19    else
20      var bit: bv10 := if arr[index] then 1 as bv10 else 0 as bv10;
21      (bit << index) + ArrayToBv10Helper(arr, index - 1)
22  }
23 method ArrayToSequence(arr: array<bool>) returns (res: seq<bool>) // Converts boolean
24   array to boolean sequence
25   ensures |res| == arr.Length
26   ensures forall k :: 0 <= k < arr.Length ==> res[k] == arr[k]
27   {
28     res := [];
29     var i := 0;
30     while i < arr.Length
31       invariant 0 <= i <= arr.Length
32       invariant |res| == i
33       invariant forall k :: 0 <= k < i ==> res[k] == arr[k]
34       {
35         res := res + [arr[i]];
36         i := i + 1;
37       }
38   }
39 function isBitSet(x: bv10, bitIndex: nat): bool
40   requires bitIndex < 10
41   ensures isBitSet(x, bitIndex) <==> (x & (1 << bitIndex)) != 0
42   {
43     (x & (1 << bitIndex)) != 0
44   }
45
46 function Bv10ToSeq(x: bv10): seq<bool> // Converts bitvector to boolean sequence
47   ensures |Bv10ToSeq(x)| == 10
48   ensures forall i: nat :: 0 <= i < 10 ==> Bv10ToSeq(x)[i] == isBitSet(x, i)
49   {
50     [isBitSet(x, 0), isBitSet(x, 1), isBitSet(x, 2), isBitSet(x, 3),
51      isBitSet(x, 4), isBitSet(x, 5), isBitSet(x, 6), isBitSet(x, 7),
52      isBitSet(x, 8), isBitSet(x, 9)]
53   }

```

```

54
55 function BoolToInt(a: bool): int {
56     if a then 1 else 0
57 }
58
59 function XOR(a: bool, b: bool): bool {
60     (a || b) && !(a && b)
61 }
62
63 function BitAddition(s: array<bool>, t: array<bool>): seq<bool> // Performs traditional
    bit addition
64     reads s
65     reads t
66     requires s.Length == 10 && t.Length == 10
67 {
68     var a: bv10 := ArrayToBv10(s);
69     var b: bv10 := ArrayToBv10(t);
70     var c: bv10 := a + b;
71
72     Bv10ToSeq(c)
73 }
74
75 method f(s: array<bool>, t: array<bool>) returns (sresult: seq<bool>) // Generated program
    for bit addition
76     requires s.Length == 10 && t.Length == 10
77     ensures |sresult| == 10
78     ensures forall i :: 0 <= i && i < |sresult| ==> sresult[i] == ((s[i] != t[i]) != (i > 0
        && ((s[i-1] || t[i-1]) && !(sresult[i-1] && (s[i-1] != t[i-1])))))
79     ensures BitAddition(s, t) == sresult // Verification of correctness
80 {
81     var a: bool := false;
82     var b: bool := false;
83     var result: array<bool> := new bool[10];
84     var i: int := 0;
85
86     while i < result.Length
87         invariant 0 <= i <= result.Length
88         invariant forall j :: 0 <= j < i ==> result[j] == false
89     {
90         result[i] := false;
91         i := i + 1;
92     }
93
94     i := 0;
95
96     assert forall j :: 0 <= j < result.Length ==> result[j] == false;
97
98     while i < result.Length
99         invariant 0 <= i <= result.Length
100         invariant b == (i > 0 && ((s[i-1] || t[i-1]) && !(result[i-1] && (s[i-1] != t[i-1]))))
101         invariant forall j :: 0 <= j < i ==> result[j] == ((s[j] != t[j]) != (j > 0 && ((s[j
            -1] || t[j-1]) && !(result[j-1] && (s[j-1] != t[j-1])))))
102     {
103         assert b == (i > 0 && ((s[i-1] || t[i-1]) && !(result[i-1] && (s[i-1] != t[i-1]))));
104
105         result[i] := XOR(b, XOR(s[i], t[i]));
106         b := BoolToInt(b) + BoolToInt(s[i]) + BoolToInt(t[i]) > 1;
107         assert b == ((s[i] || t[i]) && !(result[i] && (s[i] != t[i])));
108
109         i := i + 1;
110     }
111
112     sresult := ArrayToSequence(result);
113 }

```


Table 2. AutoML architecture search results. All networks achieved 100% accuracy on at least one test batch.

Task #	Task Name	n	w_f	d_f	w_g	d_g	Train Loss	Test Loss
1	Binary_Addition	2	1	1	4	2	0	0
2	Base_3_Addition	2	1	1	5	2	0	0
3	Base_4_Addition	2	1	1	5	2	0	0
4	Base_5_Addition	2	1	1	5	2	0	0
5	Base_6_Addition	2	1	1	6	2	2.45e-09	2.53e-09
6	Base_7_Addition	2	1	1	10	2	2.32e-06	2.31e-06
7	Bitwise_Xor	1	1	1	2	2	0	0
8	Bitwise_Or	1	1	1	1	1	3.03e-02	3.03e-02
9	Bitwise_And	1	1	1	1	1	3.03e-02	3.03e-02
10	Bitwise_Not	1	1	1	1	1	0	0
11	Parity_Last2	1	1	1	229	2	1.68e-02	1.69e-02
12	Parity_Last3	2	1	1	5	2	1.62e-04	1.64e-04
13	Parity_Last4	3	1	1	29	2	3.07e-07	2.99e-07
14	Parity_All	1	1	1	2	2	0	0
15	Parity_Zeros	1	1	1	2	2	0	0
16	Evens_Counter	4	1	1	73	3	8.89e-05	8.88e-05
17	Sum_All	1	1	1	1	1	6.09e-08	6.13e-08
18	Sum_Last2	2	1	1	1	1	0	0
19	Sum_Last3	3	1	1	1	1	6.34e-07	6.35e-07
20	Sum_Last4	4	1	1	1	1	2.10e-04	2.11e-04
21	Sum_Last5	5	1	1	1	1	8.86e-03	8.87e-03
22	Sum_Last6	6	1	1	1	1	1.82e-02	1.81e-02
23	Sum_Last7	7	1	1	1	1	3.03e-02	3.01e-02
24	Current_Number	1	1	1	1	1	0	0
25	Prev1	2	1	1	1	1	0	0
26	Prev2	3	1	1	1	1	0	0
27	Prev3	4	1	1	1	1	0	0
28	Prev4	5	1	1	1	1	2.04e-07	2.05e-07
29	Prev5	6	1	1	1	1	6.00e-05	5.96e-05
30	Previous_Equals_Current	2	1	1	5	2	6.72e-05	6.61e-05
31	Diff_Last2	2	1	1	1	1	0	0
32	Abs_Diff	2	2	2	1	1	1.84e-07	1.84e-07
33	Abs_Current	1	1	1	2	2	4.51e-08	5.71e-08
34	Diff_Abs_Values	2	1	1	4	2	3.15e-06	2.96e-06
35	Min_Seen	1	1	1	2	2	0	0
36	Max_Seen	1	1	1	2	2	1.46e-12	0
37	Majority_0_1	1	1	1	63	2	4.03e-03	4.05e-03
38	Majority_0_2	4	1	1	98	2	1.64e-04	1.71e-04
39	Majority_0_3	21	1	1	132	3	6.94e-05	6.86e-05
40	Evens_Detector	5	1	1	163	2	8.18e-04	8.32e-04
41	Perfect_Square_Detector	48	1	1	100	2	1.92e-03	1.97e-03
42	Bit_Palindrome	18	1	1	86	2	3.81e-05	3.69e-05
43	Balanced_Parenthesis	1	1	1	16	2	7.44e-03	7.10e-03
44	Parity_Bits_Mod2	1	1	1	1	1	0	0
45	Alternating_Last3	2	1	1	3	2	1.85e-02	1.87e-02
46	Alternating_Last4	2	1	1	3	2	8.24e-06	8.09e-06
47	Bit_Shift_Right	2	1	1	1	1	0	0
48	Bit_Dot_Prod_Mod2	1	1	1	3	2	0	0
49	Div_3	2	1	1	59	2	6.40e-03	6.43e-03
50	Div_5	4	1	1	76	2	1.50e-04	1.55e-04
51	Div_7	4	1	1	103	2	6.65e-04	6.63e-04
52	Add_Mod_3	1	1	1	149	2	1.02e-03	1.04e-03
53	Add_Mod_4	2	1	1	33	2	1.53e-04	1.44e-04
54	Add_Mod_5	3	1	1	43	2	1.02e-03	1.03e-03
55	Add_Mod_6	4	1	1	108	2	6.14e-04	6.12e-04
56	Add_Mod_7	4	1	1	199	2	3.96e-04	4.07e-04
57	Add_Mod_8	67	1	1	134	2	8.53e-04	8.34e-04
58	Dithering	81	1	1	166	2	7.72e-04	7.75e-04
59	Newton_Freebody	2	1	1	1	1	2.61e-07	2.62e-07
60	Newton_Gravity	2	1	1	1	1	1.81e-07	1.87e-07
61	Newton_Spring	2	1	1	1	1	0	0
62	Newton_Magnetic	4	1	1	1	1	8.59e-05	8.60e-05