

Improving Code Generation by Training with Natural Language Feedback

Angelica Chen¹ J  r  my Scheurer^{1,2} Tomasz Korbak^{1,2,3} Jon Ander Campos^{1,4} Jun Shern Chan^{1,2}
 Samuel R. Bowman¹ Kyunghyun Cho^{1,5,6} Ethan Perez^{1,2,7}

Abstract

The potential for pre-trained large language models (LLMs) to use natural language feedback at inference time has been an exciting recent development. We build upon this observation by formalizing an algorithm for learning from natural language feedback at training time instead, which we call Imitation learning from Language Feedback (ILF). ILF requires only a small amount of human-written feedback during training and does not require the same feedback at test time, making it both user-friendly and sample-efficient. We further show that ILF can be seen as a form of minimizing the KL divergence to the ground truth distribution and demonstrate a proof-of-concept on a neural program synthesis task. We use ILF to improve a CODEGEN-MONO 6.1B model’s pass@1 rate by 38% relative (and 10% absolute) on the Mostly Basic Python Problems (MBPP) benchmark, outperforming both fine-tuning on MBPP and fine-tuning on repaired programs written by humans. Overall, our results suggest that learning from human-written natural language feedback is both more effective and sample-efficient than training exclusively on demonstrations for improving an LLM’s performance on code generation tasks.

1. Introduction

An important task for the field of software engineering is program synthesis, the automatic generation of computer programs from an input specification (*e.g.* a natural language task description or a set of input-output examples) (Manna & Waldinger, 1971). Effective program synthesis can not only improve the efficiency of software developers (Ziegler et al., 2022), but also increase the accessibility of

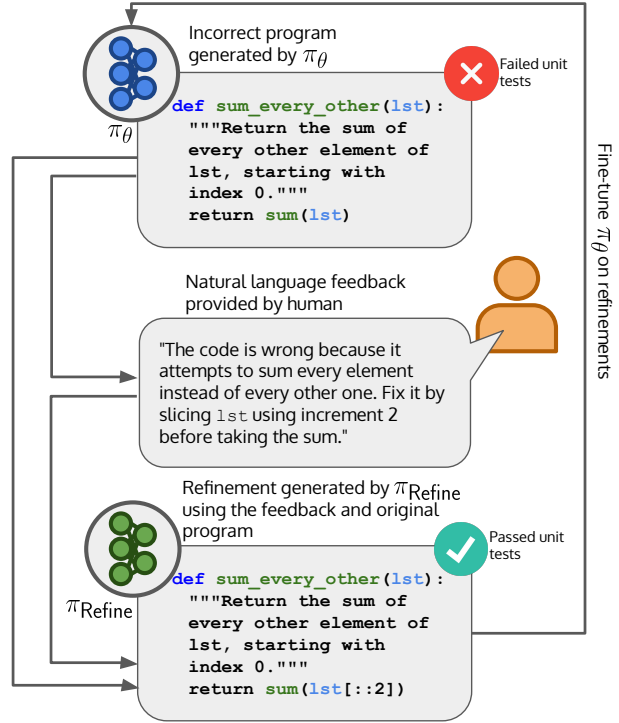


Figure 1. An overview of imitation learning from language feedback (ILF) for code generation. Given an initial LLM π_θ , we sample programs from π_θ that do not pass unit tests (indicated by the red X). Human annotators write natural language feedback for the incorrect program and a model π_{Refine} generates a *refinement* - *i.e.* an improved version of the original program that incorporates the feedback and passes the unit tests (indicated by the green checkmark). Finally, we fine-tune π_θ on the refinements.

writing code in general. Recently, pre-trained large language models (LLMs) have demonstrated impressive success on program synthesis (Chen et al., 2021; Li et al., 2022; Austin et al., 2021; Nijkamp et al., 2022; Xu et al., 2022, *inter alia*) but still struggle to consistently generate correct code, even with large-scale pre-training (Chen et al., 2021).

We hypothesize that these failures can be largely attributed to modern LLM pre-training set-ups. For instance, code pre-training datasets consist mostly of unfiltered code scraped

¹New York University ²FAR AI ³University of Sussex ⁴HiTZ Center, University of the Basque Country UP-V/EHU ⁵Genentech ⁶CIFAR LMB ⁷Anthropic. Correspondence to: Angelica Chen <angelica.chen@nyu.edu>, Ethan Perez <ethan@anthropic.com>.

from the Internet, which contains a significant number of security vulnerabilities (Kang et al., 2022) and bugs (Chen et al., 2021). This training signal also consists exclusively of offline demonstrations, without any signal from trial-and-error or interactive guidance that penalizes the model’s buggy outputs. As such, we hypothesize that supervising LLMs with explicit human-written feedback on the model’s own outputs can be more effective at training models to produce functionally correct code.

In particular, an intuitive and rich form of feedback to provide to LLMs is *natural language* feedback. We argue that LLMs are naturally able to incorporate written feedback, which has been shown to significantly improve a code generation model’s pass rates when the feedback is provided at test time (Nijkamp et al., 2022; Austin et al., 2021). In our work, we build upon this observation by exploring the use of natural language feedback during the training process itself, rather than just during inference. We conjecture that such feedback provides expressive and targeted information about a code generation model’s current failings in a sample-efficient manner. More broadly, this approach also represents a weak version of *scalable oversight* (Bowman et al., 2022), in that model overseers can improve a model merely by evaluating its outputs, without manually generating new demonstrations, in a way that takes advantage of the capabilities that are being supervised.

To train LLMs with language feedback, we propose an algorithm called Imitation learning from Language Feedback (ILF; Algorithm 1), which extends the work of Scheurer et al. (2022), who study the impact of learning from language feedback on text summarization models. Scheurer et al. (2022) improves a summarization model by training the base model on improved summaries generated from the model’s original summaries and human-written feedback. Our work builds upon Scheurer et al. (2022) in a number of ways: (1) by formalizing the algorithm and generalizing it into a form that can be applied to any task (our ILF algorithm in Section 2.2), (2) by detailing how the reward function can be adapted for code generation, and (3) by demonstrating a proof-of-concept of ILF for code generation.¹ ILF improves the correctness of programs generated by a baseline code generation model π_θ by training a separate model π_{Refine} to use language feedback to repair the incorrect π_θ -generated programs. (We refer to the repaired programs as *refinements*.) We then improve π_θ by fine-tuning it on the π_{Refine} -generated refinements that pass unit tests, yielding a final improved model π_{θ^*} . This procedure may be run iteratively to continue improving the model, which we show can be seen as minimizing the expected KL divergence from a target ground truth distribution (Section 2).

¹We open-source our code and annotated data at <https://github.com/nyu-ml1/ILF-for-code-generation>.

We demonstrate a proof of concept of ILF for code generation by showing that it improves a CODEGEN-MONO 6.1B model’s pass@1 rate on the Mostly Basic Python Problems (MBPP) benchmark (Odena et al., 2021) by 38% relative (10% absolute) over its zero-shot performance. It also outperforms fine-tuning on the MBPP-provided code by 64% (14% absolute, see Section 3.2). We further find that the refinements generated during ILF do indeed leverage the human-written feedback (Section 3.1) – when the feedback is unhelpful or irrelevant, we observe steep drops in code correctness. The quality of the feedback is also crucial – LLM-generated feedback yields far lower final pass rates than human-written feedback (Section 3.3). Despite the success of our approach, we still observe concrete limitations – for instance, π_{Refine} is less effective at incorporating feedback when the feedback addresses multiple bugs (Section 3.5), which suggests headroom for future work or more capable LLMs to base π_{Refine} on. Overall, our results – as well as our additional results on text summarization, using a similar technique in Scheurer et al. (2023) – suggest that human-written feedback is a powerful, information-rich form of supervision for LLMs.

2. Method

2.1. Preliminaries

Here, we formally describe the problem we aim to tackle, before introducing our algorithm. Suppose we start with vocabulary \mathcal{V} and a pre-trained language model π_θ parameterized by θ . $\pi_\theta : \mathcal{V}^* \rightarrow [0, 1]$ is a probability distribution over sequences of tokens $x \in \mathcal{V}^*$, where \mathcal{V}^* is the Kleene closure of \mathcal{V} . We also have a dataset of tasks $\mathcal{D} = \{(t, u)\}$. A task (t, u) consists of a task description $t \in \mathcal{T}$ (e.g. “Write a function that computes the prime factorization of an input integer.”) and a suite $u = \text{UNITTESTS}(t) \in \mathcal{U}$ of unit tests associated with task t . Finally, let $\text{EVAL} : \mathcal{V}^* \times \mathcal{T} \rightarrow \{0, 1\}$ be a unit test verification function that indicates whether a program $x \sim \pi_\theta(\cdot | t)$ passes all the unit tests in $\text{UNITTESTS}(t)$:

$$\text{EVAL}(x, t) := \begin{cases} 1, & \text{if } x \text{ passes test suite } \text{UNITTESTS}(t), \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

We also define a fine-tuning function $\text{FINETUNE}(\pi_\theta, \mathcal{D})$ that applies a gradient-based optimization algorithm to π_θ using the associated loss objective calculated over dataset \mathcal{D} .

2.2. Imitation Learning From Language Feedback

Our goal is to sample a diverse set of high-quality programs $x_1 \sim \pi_\theta(\cdot | t)$ for any given task t sampled from the task distribution $p(t)$. We do so by fitting an auto-regressive LLM π_θ to approximate a ground truth distribution $\pi_t^*(x_1)$ that assigns a probability to x_1 that is proportional to its

Algorithm 1 Imitation learning from natural language feedback for code generation.

-
- 1: **Input:** Dataset \mathcal{D} , initial LLM π_θ , unit test verification function EVAL, LLM $\pi_{\text{Refine}} : \mathcal{V}^* \rightarrow [0, 1]$ trained to incorporate feedback into code
 - 2: $C \leftarrow \{(x_0, t, u) \mid x_0 \sim \pi_{\theta_k}(\cdot|t), \text{EVAL}(x_0, t) = 0, (t, u) \in \mathcal{D}\}$
 - 3: $C_{\text{annotated}} \leftarrow \{(x_0, f, t) \mid (x_0, t, u) \in C\}$ \triangleright Humans write feedback f for $x_0 \in C$.
 - 4: $R \leftarrow \{(t, x_1) \sim \pi_{\text{Refine}}(\cdot|t, x_0, f) \mid \text{EVAL}(x_1, t) = 1, (x_0, f, t) \in C_{\text{annotated}}\}$ $\triangleright \pi_{\text{Refine}}$ generates refinements x_1 that incorporate feedback f into x_0 .
 - 5: $\pi_{\theta^*} \leftarrow \text{FINETUNE}(\pi_\theta, R)$
-

quality, as measured by a reward function R . Fitting π_θ to approximate π_t^* can be seen as minimizing the expected KL divergence from π_t^* to π_θ over the task distribution $p(t)$:

$$\min_{\theta} \mathbb{E}_{t \sim p(t)} [\text{KL}(\pi_t^*, \pi_\theta(\cdot|t))] \quad (2)$$

where

$$\pi_t^*(x_1) \propto \exp(\beta R(x_1, t)) \quad (3)$$

In this work we use the unit test verification function EVAL directly as our reward function R , but R can also be a function of any number of other signals, such as stack traces or compiler outputs.

Minimizing the objective in Equation 2 is equivalent to supervised learning, *i.e.* minimizing the cross-entropy loss:

$$\mathcal{L}(\theta) = - \mathbb{E}_{t \sim p(t)} [\mathcal{L}_\theta(t)], \quad (4)$$

where

$$\mathcal{L}_\theta(t) = \sum_{x_1} \pi_t^*(x_1) \log \pi_\theta(x_1|t). \quad (5)$$

Rather than computing this loss over the exponentially large space of all possible x_1 's, we instead use Monte-Carlo sampling over a small set of x_1 's drawn from π_t^* . However, this is still intractable because we cannot sample directly from π_t^* . Instead, we approximate π_t^* using importance sampling with a proposal distribution $q_t(x_1)$:

$$\mathcal{L}_\theta(t) = \sum_{x_1} q_t(x_1) \frac{\pi_t^*(x_1)}{q_t(x_1)} \log \pi_\theta(x_1|t) \quad (6)$$

which assigns higher weights to higher quality programs x_1 .

2.3. Proposal Distribution q

Intuitively, we aim to design q_t to be as close as possible to π_t^* , which we accomplish by incorporating pieces of natural language feedback f that give information about how to transform a low-reward program x_0 into a higher-reward program x_1 . This can be achieved by (i) identifying a program $x_0 \sim \pi_\theta(\cdot|t)$ that does not currently pass the test suite (*i.e.* $\text{EVAL}(x_0, t) = 0$), (ii) asking for natural language feedback f about bugs in x_0 , (iii) using f to transform the original program x_0 into a *refinement* x_1 that incorporates

the feedback and passes the test suite (*i.e.* $\text{EVAL}(x_1, t) = 1$), and (iv) assigning higher weight to x_1 .

We can formalize this procedure as follows. Let $\pi_\psi(x_1|t, x_0, f)$ be a distribution over programs x_1 that improve x_0 by incorporating the feedback f and $p_{\mathcal{F}}(f|t, x_0, \text{EVAL}(x_0, t) = 0)$ be the distribution of pieces of feedback f for incorrect program x_0 and task t . We can then define our proposal distribution as:

$$\begin{aligned} q_t(x_1) = \sum_{x_0, f} & \pi_\theta(x_0|t) \times \delta_0(\text{EVAL}(x_0, t) | x_0, t) \\ & \times p_{\mathcal{F}}(f|t, x_0, \text{EVAL}(x_0, t) = 0) \\ & \times \pi_\psi(x_1|t, x_0, f) \\ & \times \delta_1(\text{EVAL}(x_1, t) | t, x_1), \end{aligned} \quad (7)$$

where δ_0 and δ_1 are the Dirac delta distributions centered at 0 and 1, respectively. Then this proposal distribution is guaranteed to place higher probability mass on higher-quality programs (in terms of unit test pass rate) than π_θ since the term $\delta_1(\text{EVAL}(x_1, t) | t, x_1)$ equals 0 for incorrect programs x_1 .

We approximate sampling from q by considering each of the terms in Equation 7 in order:

1. We first sample from $\pi_\theta(x_0|t) \times \delta_0(\text{EVAL}(x_0, t) | x_0, t)$ by rejection sampling from π_θ . In other words, we sample programs x_0 from π_θ for task t and only keep those that fail the test suite (*i.e.* $\text{EVAL}(x_0, t) = 0$; step 2 of Algorithm 1).
2. We approximate sampling from $p_{\mathcal{F}}(f|t, x_0, \text{EVAL}(x_0, t) = 0)$ by having humans annotate programs x_0 (paired with their corresponding task descriptions t and test suites u) with natural language feedback (step 3 of Algorithm 1).
3. We approximate sampling from $\pi_\psi(x_1|t, x_0, f)$ by sampling from π_{Refine} , a model capable of generating refinements given the task description, original programs, and human-written feedback.
4. Finally, the term $\delta_1(\text{EVAL}(x_1, t) | t, x_1)$ corresponds to another filter: we only keep refined programs x_1 that pass the test suite.

Next, we consider more concrete details of how this sampling is accomplished.

Training π_{Refine} ILF assumes the availability of feedback but not necessarily of the repaired code/refinements, for a variety of reasons. We assume that program synthesis may be a task for which writing high-level natural language feedback is often less laborious than performing program repair. Although writing feedback involves identifying at a high level what is wrong with the program and how it should be fixed, program repair may involve the additional steps of refactoring, looking through documentation, and testing. Moreover, past work (Austin et al., 2021; Nijkamp et al., 2022) has indicated that certain large LLMs can proficiently incorporate the feedback at inference time, assuming access to accurate and high-quality feedback. As such, ILF assumes access to some model π_{Refine} that is capable of producing a refinement given the original program and feedback.

π_{Refine} can take a variety of forms, but we fine-tune a pre-trained CODEGEN-MONO 6.1B model as our π_{Refine} . We create a training dataset for π_{Refine} by further annotating a subset of $C_{\text{annotated}}$ with refinements x_1 that repair incorrect programs x_0 by incorporating feedback f , such that $\text{EVAL}(x_1, t) = 1$ for $(x_0, f, t) \in C_{\text{annotated}}$. Further details of our dataset and annotation procedure are in Section 3.

3. Experiments and Results

Having described our high-level approach, we now explain the experimental setup we use to test ILF.

Dataset We train and evaluate our models on the Mostly Basic Python Problems (MBPP) dataset (Odena et al., 2021). MBPP contains 974 Python programming tasks designed to be solvable by entry-level coders. Each task contains a natural language task description t (e.g., “Write a function to return the prime factorization of the input.”), a gold solution, and a suite u of three unit tests. Since the task descriptions are sometimes ambiguous, we include one unit test in the task description. The addition of the unit test helps to specify the input and output format of each task. We hold out the remaining unit tests for the evaluation of our generated programs.

MBPP includes a designated prompt/training/validation/test split of the dataset, but we re-split the dataset into the following splits:

- **MBPP_{Refine}**: These are tasks with IDs in the range 111-310 for which CODEGEN-MONO 6.1B did not generate any correct completions. This split is used to train π_{Refine} .
- **MBPP_{Train}**: These are tasks with IDs in the range 311-974 for which CODEGEN-MONO 6.1B did not gener-

ate any correct completions. This split is first used to evaluate the correctness of refinements generated by π_{Refine} . Then, the correct refinements in this split are used to train π_{θ} to obtain π_{θ^*} (step 5 in Algorithm 1).

- **MBPP_{Test}**: These are tasks with IDs in the range 11-110 that we use to evaluate the final performance of π_{θ^*} . Unlike the previous two splits, we use *all* tasks in this split, rather than only the tasks for which CODEGEN-MONO 6.1B did not originally generate correct programs for. This allows us to better compare the baseline performance of π_{θ} with that of π_{θ^*} .

We use this modified split so that a larger portion of the dataset can be used to train the final model π_{θ^*} , whereas smaller portions are allocated for training π_{Refine} and evaluating π_{θ^*} . We do not make use of the prompt split (IDs 1-10).

Models Throughout this paper, we use a pre-trained CODEGEN-MONO 6.1B model (Nijkamp et al., 2022) as our π_{θ} . It is pre-trained sequentially on THEPILE (Gao et al., 2020), BIGQUERY (Nijkamp et al., 2022), and BIGPYTHON (Nijkamp et al., 2022). We selected this model because it is open-source, can be fine-tuned on a single 4×100 A100 (80 GB) node, and demonstrated pass@k scores comparable to CODEX-12B (Chen et al., 2021; Nijkamp et al., 2022).

To implement our algorithm, we independently fine-tune two separate instances of CODEGEN-MONO 6.1B to create π_{Refine} and the final model π_{θ^*} . We train π_{Refine} using pairs of incorrect programs and human-written feedback as inputs, with human-written refinements as targets (using the format in Figure 2). In contrast, we train π_{θ^*} using natural language task descriptions from MBPP as the inputs and π_{Refine} -generated refinements as the targets. Further training details are in Appendix A.1.

Evaluation We evaluate all code generations in this paper using the *pass@k* metric introduced in Kulal et al. (2019). It estimates the rate for which ≥ 1 of k model samples passes all the unit tests. We use the empirical estimate of this quantity from Chen et al. (2021), an unbiased estimator given by:

$$\text{pass@k} := \mathbb{E}_{\text{task}} \left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \right] \quad (8)$$

for n total programs (where $n \geq k$) and c correct programs for the given task.

Human Annotation We hire annotators via Surge AI² to write both natural language feedback and refinements for

²www.surgehq.ai

Prompt	Expected completion
<p>OLD CODE:</p> <pre>""" Write a python function to find the sum of the three lowest positive numbers from a given list of numbers. >>> Example: sum_three_smallest_nums([10,20,30, 40,50,60,7]) = 37 """ def sum_three_smallest_nums(lst): lst.sort() return sum(lst[:3])</pre> <p>FEEDBACK:</p> <p>This code finds the sum of the smallest 3 numbers, not the smallest 3 positive numbers. It needs to disregard negatives and 0.</p> <p>REFINEMENT:</p>	<pre>""" Write a python function to find the sum of the three lowest positive numbers from a given list of numbers. >>> Example: sum_three_smallest_nums([10,20,30, 40,50,60,7]) = 37 """ def sum_three_smallest_nums(lst): lst = [x for x in lst if x > 0] lst.sort() return sum(lst[:3])</pre>

Figure 2. An example of a zero-shot LLM prompt for repairing incorrect code based on human-written feedback.

incorrect programs generated by CODEGEN-MONO 6.1B. For each task that CODEGEN-MONO 6.1B generated no correct programs for, we ask the workers to first select one of the incorrect programs to write feedback and refinement for. We specify that the workers should select a sample that seems relatively easy to correct (*i.e.* could be minimally corrected to pass the unit tests). Then, they are asked to write feedback that describes what is wrong with the current code and how to fix it. For the refinement, they are asked to copy over the original code and make the *minimum number of edits necessary* to incorporate the feedback and pass all the unit tests. The full set of worker instructions can be found in Appendix A.2.

We keep all annotations for which the refinement passes all tests in the task’s test suite, the feedback is correct (as manually verified by the authors), and the Levenshtein edit distance between the refinement and the original program is less than 50% of $\max(\text{len}(\text{refinement}), \text{len}(\text{original program}))$. The final dataset consists of 195 triples of (incorrect program, human-written feedback, human-written refinement). On average, workers are paid \$23 per annotated sample and take 27 minutes/sample, with a 10th percentile of 4 minutes and a 90th percentile of 43 minutes.

Although the ILF algorithm only requires the collection of human-written feedback for the tasks in MBPP_{Train} (assuming access to some π_{Refine} that is already fine-tuned or can generate refinements via few-shot prompting), we collect both human-written feedback and refinement for all splits of the data so that we can conduct further analyses of

our method. For instance, this allows us to compare fine-tuning on π_{Refine} -generated refinements with fine-tuning on human-written refinements. When scaled to other pairs of model and task, ILF requires new feedback annotations, but it is possible that using ILF on one dataset will improve the model’s abilities on another dataset for a similar task. We leave analyses of scaling ILF across different tasks and models to future work.

Table 1. Initial zero-shot CODEGEN-MONO 6.1B performance on the entire MBPP dataset. “1+ Correct” refers to the percentage of tasks for which CODEGEN-MONO 6.1B generated at least one program that passed all unit tests.

Metric	Zero-Shot CODEGEN-MONO 6.1B
Pass@1	31%
Pass@10	63%
1+ Correct	67%

Table 2. Evaluations of 1-shot refinements generated by CODEGEN-MONO 6.1B (before ILF) given either related or unrelated text feedback in the prompt. Feedback is provided only for tasks on which CODEGEN-MONO 6.1B previously did not output any correct programs.

Prompt Type	CODEGEN-MONO 6.1B	
	Pass@1 ↑	Pass@10 ↑
Code + feedback	2.0%	13.8%
Code + unrelated feedback	0.4%	4.0%

3.1. CODEGEN-MONO 6.1B Incorporates Feedback

We first verify that our baseline model can use feedback to repair incorrect code, a pre-requisite for ILF to work. We evaluate CODEGEN-MONO 6.1B’s ability to generate refinements given pairs of (incorrect code, natural language feedback), both in a few-shot manner and after fine-tuning. Feedback is only required for tasks for which π_θ is initially unable to produce a correct response, so we first evaluate CODEGEN-MONO 6.1B zero-shot on all of MBPP, generating 30 programs per task with temperature 0.8. Table 1 shows the resulting pass rates. There were 321 tasks for which zero-shot CODEGEN-MONO 6.1B yielded no correct samples (from Table 1: $(100\% - 67\%) \times 974 \text{ tasks} \approx 321$). We then annotate one incorrect program per task with both feedback and refinement, as described in Section 3.

Few-Shot Feedback Incorporation We use the human feedback annotations to create few-shot feedback prompts, formatted as in Figure 2. We evaluate CODEGEN-MONO 6.1B’s ability to produce refinements that incorporate the feedback and pass the unit tests. However, producing a refinement that passes the unit tests does not guarantee that the feedback has been incorporated; there can be multiple solutions to a programming task, including ones that are functional but completely different and not using the feedback to improve upon the original code. Alternatively, the model may already be able to repair programs without feedback. Thus, we also evaluate the pass rate after shuffling the feedback samples in the dataset, to evaluate if the model’s ability to repair code degrades when presented with unrelated feedback.

The results are shown in Table 2. CODEGEN-MONO 6.1B’s ability to incorporate relevant feedback on this particular set of program is low, with pass@10 reaching only 13.8%. However, the gap in accuracy between CODEGEN-MONO 6.1B-generated refinements on relevant versus irrelevant feedback is significant, with pass@10 decreasing by 71% (relative; $13.8\% \rightarrow 4.0\%$), indicating that the model is indeed using the feedback.

Training π_{Refine} Next, we examine whether we can improve our ability to repair programs given feedback by fine-tuning a separate model specifically to perform this task. Our training examples consist of triples of incorrect program, human-written feedback, and human-written refinement. We train the model to maximize the likelihood of the refinement given the program and feedback. The incorrect programs were generated by CODEGEN-MONO 6.1B zero-shot on MBPP tasks, and the feedback and refinements were written by human annotators, as discussed in Section 3. We only included tasks for which none of CODEGEN-MONO 6.1B’s generated programs were correct, yielding 44 tasks in the training dataset (forming the

split MBPP_{Refine}) and 128 tasks in the evaluation dataset (forming the split MBPP_{Train}). We asked human annotators to write refinements of the original code that incorporated their own previously written feedback, passed the unit tests, and made only minimal edits to the code (see Section 3). The format of the training data also matched the few-shot prompt format (Figure 2) but without the in-context examples of refinements. We denote this model as π_{Refine} , as described in Section 2.3.

Table 3. Pass rates of π_{Refine} -generated refinements versus zero-shot CODEGEN-MONO 6.1B programs for tasks in MBPP_{Train}.

Metric	π_{Refine}	Zero-shot CODEGEN-MONO 6.1B
Pass@1	19%	0%
Pass@10	47%	0%
1+ correct	61%	0%

Table 3 shows the pass rates for π_{Refine} on the evaluation dataset, which were produced by sampling 30 refinements per task with temperature 0.8. Fine-tuning significantly improves CODEGEN-MONO 6.1B’s ability to incorporate feedback compared to 1-shot refinement, increasing pass rates more than three-fold ($2 \rightarrow 19\%$ pass@1, $13.8 \rightarrow 47\%$ pass@10, from Tables 2 and 3). Furthermore, 61% of tasks had at least one correct refinement. This is particularly significant when considering the fact that we selected only tasks for which a non-finetuned CODEGEN-MONO 6.1B model did not originally output any correct programs for (the rightmost column in Table 3). For the 61% of validation tasks that π_{Refine} generated a correct refinement for, we randomly selected one such correct program for each task to form the training dataset for our final model π_{θ^*} , yielding a final training dataset of 78 examples.

3.2. ILF Yields Pass Rates Higher Than Fine-Tuning on Gold Data or Human-Written Programs Alone

Given that our refinements improve over the initial programs, we now fine-tune on the refinements to improve our code generation model. As discussed earlier, we use the correct refinements (as evaluated by the unit tests) that π_{Refine} generated for its evaluation dataset as the training dataset for π_{θ^*} . Since π_{θ^*} is meant to generate code from a natural language task description (rather than to incorporate feedback into a refinement), the inputs of our training dataset are the MBPP prompts and the targets are the 78 π_{Refine} -generated refinements described in the previous section. We also compare the performance of π_{θ^*} against that of CODEGEN-MONO 6.1B evaluated in a zero-shot manner, CODEGEN-MONO 6.1B fine-tuned on the gold programs from the MBPP dataset, and CODEGEN-MONO 6.1B fine-tuned on our human-written refinements. For all fine-tuning experiments, we train on programs corresponding to the

Table 4. Final performance of π_{θ^*} on MBPP_{Test}, compared to other ablations and baselines. All results are calculated using 30 output samples with temperature 0.8. All the methods are built on the CODEGEN-MONO 6.1B model.

Method	Feedback Source	Fine-Tuning Data	Pass Rates of π_{θ^*}	
			Pass@1	Pass@10
ILF	Humans	π_{Refine} Refinements	36%	68%
Ablations	1-shot InstructGPT	1-shot InstructGPT Refinements	19%	55%
	2-shot InstructGPT	2-shot InstructGPT Refinements	25%	59%
Gold Standards	-	MBPP Gold	22%	63%
	-	Human Refinements	33%	68%
Baseline (zero-shot)	-	-	26%	59%

same set of task IDs as the ones used in π_{θ^*} ’s training dataset.

Additionally, we evaluate the impact of ablating the human annotations in our algorithm by using an LLM in place of humans to generate the feedback and refinements (replacing steps 3 and 4 in Algorithm 1). For the LLM, we use GPT-3.5 fine-tuned with Feedback Made Easy (FeedME; text-davinci-002 on the OpenAI API)³. We refer to this model as InstructGPT, which is the series of OpenAI models that FeedME belongs to (OpenAI, 2022). We use InstructGPT to generate both the feedback and refinements on the original programs. We then fine-tune CODEGEN-MONO 6.1B on the model-generated refinements.

The results of our ILF algorithm compared to the baselines and ablations are shown in Table 4. ILF yields the highest pass@1 and pass@10 rates, despite how few samples of feedback and refinements we use. The pass@1 rate in particular shows a significant increase in improvement over the zero-shot baseline, representing a 10% absolute increase (38% relative increase). Pass@1 improvements are especially helpful for assisting with software engineering, where it is more helpful to suggest a single correct completion rather than 10 possible completions for the user to select from.

Compared to the gold standards, ILF outperforms both fine-tuning on MBPP gold programs and human-written refinements on the pass@1 metric, yielding 14% absolute (64% relative) and 3% absolute (9% relative) increases in pass@1 rates, respectively. However, training on human-written refinements yielded comparable pass@10 rates as ILF, which is unsurprising since π_{Refine} was trained on human-written refinements. When human-written feedback and π_{Refine} -generated refinements are ablated (the “Ablations” section of Table 4), ILF also outperforms training on both 1-shot and 2-shot InstructGPT-generated refinements by 17% and 11% absolute (89% and 44% relative), respectively.

³Details at beta.openai.com/docs/model-index-for-researchers



Figure 3. Histogram of the perplexities of the various training data sources, as measured using a pre-trained CODEGEN-MONO 6.1B model.

Analysis of Training Data Sources However, we also note the surprising fact that merely training on a small sample of the MBPP gold programs did not make a significant difference in accuracy over zero-shot inference. We speculate that the gold programs from the MBPP dataset may be somewhat out-of-distribution for CODEGEN-MONO 6.1B. To test this hypothesis, we computed the perplexity of the MBPP gold programs, the π_{Refine} -generated refinements, and the human-written refinements using the pre-trained CODEGEN-MONO 6.1B model. The results are shown in Figure 3. While the distributions of all three data sources look similar, the MBPP dataset contains more high-perplexity programs (*i.e.* programs with perplexity $\geq 10^2$) than either the π_{Refine} -generated refinements or the human-written refinements. As a result, it is likely easier for CODEGEN-MONO 6.1B to learn from the latter two datasets, since they are closer to CODEGEN-MONO 6.1B’s original distribution while still being functionally correct.

Furthermore, ILF is particularly useful for settings where large amounts of gold code are not available. In this setting, ILF can be thought of as a method of not only generating

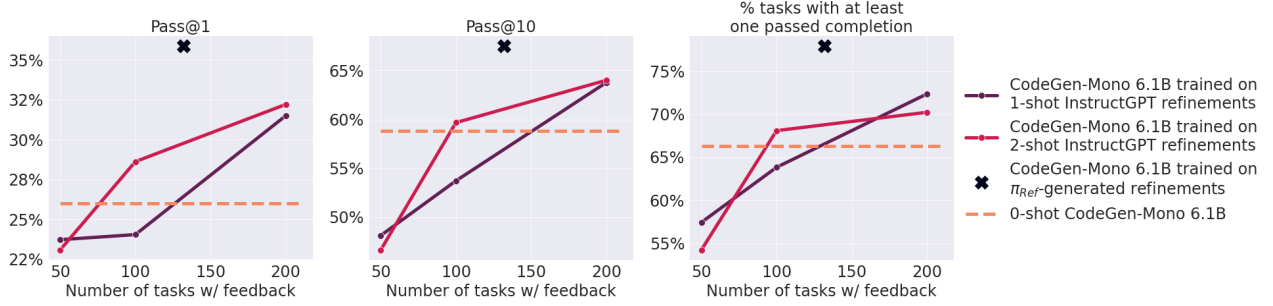


Figure 4. Training dataset size versus CODEGEN-MONO 6.1B pass rates on MBPP tasks 11-111 after fine-tuning on InstructGPT-generated refinements, versus the performance of π_{θ^*} (the model produced by our approach). \times marks the performance of π_{θ^*} , whereas the solid lines plot the performance of CODEGEN-MONO 6.1B after fine-tuning on correct refinements generated by InstructGPT, using feedback also generated by InstructGPT. The dashed line indicates the zero-shot pass rate of a pre-trained CODEGEN-MONO 6.1B model.

more training data, but training data that is closer to the model’s original outputs in data representation space and that specifically repairs the kinds of bugs that the original model generates. As a result, fine-tuning the model on π_{Refine} -generated refinements does not require adjusting the weights as much as fine-tuning the model on the MBPP gold programs would, even though both training datasets contain the same number of functionally correct programs.

3.3. Scaling Up Model Feedback Does Not Offer the Same Benefits As Human Feedback

Since high quality human feedback can be expensive to collect, we also evaluated how much model feedback might yield the same benefit as our sample of human-written feedback. To do so, we randomly select k tasks from the set of MBPP tasks for which CODEGEN-MONO 6.1B did not originally output a correct answer, and prompt InstructGPT to generate both the feedback and the refinement. We then evaluate the refinements for correctness and train CODEGEN-MONO 6.1B on the correct refinements. We use $k \in \{50, 100, 200\}$ and generate 30 output samples at temperature 0.8 for all stages of the experiment. We are limited to these k values due to the small number of tasks we have in MBPP_{Train}, but future work may investigate scaling up these experiments by using larger datasets or automatically generating new tasks and unit tests for the training dataset. Further training details are listed in Appendix A.1.

The results are shown in Figure 4. Although increasing the quantity of InstructGPT-generated feedback offers modest improvements in pass rates, these improvements do not yield pass rates as high as those of π_{θ^*} , even though π_{θ^*} uses only a total of 122 pieces of feedback throughout its training process (44 for training π_{Refine} and 78 for generating refinements to train π_{θ^*} on). However, as pre-trained large language models continue to improve dramatically

in quality, we expect that this gap between human- and model-written feedback will increasingly narrow.

Table 5. The proportion of the feedback that addressed each type of bug, for feedback sourced from humans and InstructGPT. Each sample of feedback can be tagged with multiple categories, so the quantities in each column do not necessarily add up to 100%.

Feedback Category	% of Feedback	
	Human	InstructGPT
Logic	30%	46%
Formatting	36%	14%
Missing step	10%	6%
Algebra	10%	8%
Recursion	4%	14%
Regex	6%	6%
Function semantics	2%	4%
Dynamic programming	2%	0%
Extra step	0%	12%
No feedback needed	0%	14%
Unrelated	0%	8%

Table 6. Descriptive statistics for the human- versus InstructGPT-generated feedback. The * indicates that the metric was computed on the random sample of 50 that we manually inspected, whereas the other metrics are computed from the full dataset.

	Source of Feedback	
	Human	InstructGPT
Avg. num. of bugs addressed*	1.8	1.1
Avg. num. of words	68.9 \pm 48.2	24.2 \pm 28.6

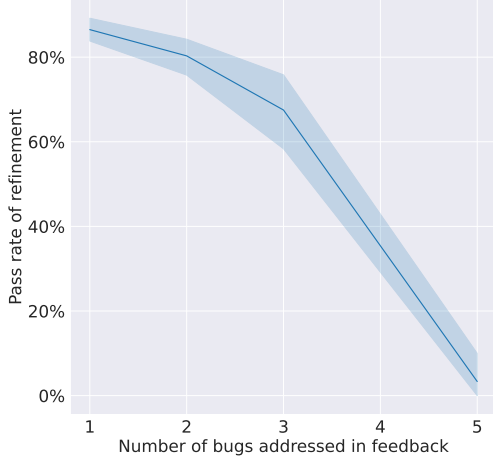


Figure 5. The number of bugs addressed in the feedback versus the pass rate of π_{Refine} ’s refinements.

3.4. Human Feedback Is More Informative Than InstructGPT Feedback

To better understand why human feedback produced greater improvements in pass rate than InstructGPT feedback, we randomly selected 50 samples of feedback for each source (*i.e.* human or InstructGPT) and annotated the number and types of bugs that each feedback sample addressed. The results are shown in Tables 5 and 6. We observed that InstructGPT often gave no feedback (*e.g.* “The code is correct” or “Great job!”), provided feedback that was irrelevant or incorrect, or restated the task description instead of addressing what should be repaired about the code. Despite this, InstructGPT’s refinements were often correct even if the feedback itself wasn’t. Human-written feedback addressed more bugs on average and never gave irrelevant feedback. We provide further examples of the differences between human and InstructGPT feedback in Appendix A.3.

3.5. π_{Refine} Struggles To Incorporate Feedback Addressing Many Bugs

Lastly, we explored whether the number of bugs addressed in the feedback affected π_{Refine} ’s ability to repair the original code sample. The results are shown in Figure 5. The greater the number of bugs addressed, the lower the average pass rate of π_{Refine} ’s refinements. This suggests that a promising direction for future work might consist of automatically decomposing the feedback into multiple steps and having π_{Refine} incorporate the feedback one step at a time. Indeed, Nijkamp et al. (2022) show that the CODEGEN models are often more effective at following instructions when the instructions are given across multiple turns, and recent Chain-of-Thought work (Wei et al., 2022) illustrates a similar prompting technique.

4. Related Work

LLMs for Program Synthesis Our work builds on a large body of literature that explores the use of pre-trained LLMs for neural program synthesis. Many general purpose LLMs, although not pre-trained specifically for code generation, have demonstrated impressive proficiency at solving code challenges since they are pre-trained on large corpora of text such as THE PILE (Gao et al., 2020) that contain a small percentage of code content (Austin et al., 2021; Wang & Komatsuzaki, 2021; Black et al., 2022; Nijkamp et al., 2022). Yet other recent LLMs for program synthesis are trained on solely source code files (Wang et al., 2021; Zan et al., 2022; Li et al., 2022; Xu et al., 2022), or on both text and source code documents – sometimes either in succession (Chen et al., 2021; Nijkamp et al., 2022; Bai et al., 2022a), in a mixed corpus (Workshop et al., 2022), or on mixed natural language-programming language documents (Feng et al., 2020).

Learning from Human Feedback Our algorithm is inspired by a number of past works that have trained models to learn from feedback. A common technique is reinforcement learning from human feedback (RLHF Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022), which trains models to satisfy human preferences. However, our algorithm is closer to works that use natural language feedback, rather than comparisons between different choices. Elgohary et al. (2020); Austin et al. (2021); Nijkamp et al. (2022) all demonstrate that code LLM performance generally improves when prompted with natural language feedback, though Nijkamp et al. (2022) observes that the feedback is more effective when it is given one step at a time. Our work differs from these in that ILF learns from the feedback at training time, not at inference time.

Bai et al. (2022a) also uses natural language feedback during the training process, but as part of an RLHF algorithm instead where the feedback is used to solicit different responses from the digital assistant, the responses are ranked by crowdworkers, and the rankings are used to train the preference model. However, they note that this form of learning from natural language feedback does not measurably improve their code generation model more than simply prompting.

Outside of program synthesis, we show in our other work (Scheurer et al., 2023) that ILF is also effective for text summarization. In addition to re-formulating the reward function $R(\cdot)$ for summarization, Scheurer et al. (2023) additionally demonstrates that an instruction-finetuned LLM can evaluate its own outputs and select the best one. Similar to our results on code generation, Scheurer et al. (2023) shows that ILF outperforms all supervised fine-tuning baselines on text summarization. This aligns with numerous other works

that have explored supervision via natural language in other ways, such as via explanations (Camburu et al., 2018; Hase & Bansal, 2021; Pruthi et al., 2021; Lampinen et al., 2022, *inter alia*) and as part of RL systems (Fidler et al., 2017; Luketina et al., 2019; Lin et al., 2020, *inter alia*).

5. Conclusion

We have shown that ILF can significantly improve the quality of a code generation model, even with just a small sample of human-written feedback and refinements. This approach is theoretically justified as minimizing the expected KL divergence between π_θ and a target ground-truth distribution, where we acquire signal from the latter via human-written natural language feedback.

This approach is also appealing because it is not model-specific (in the sense that ILF can be used with any type of base model π_θ , assuming the existence of a sufficiently capable LLM to act as π_{Refine}), and can be conducted in multiple rounds to continuously improve the model. Furthermore, it is notable that our approach generates training data that is not only correct, but targets the specific kinds of bugs that the model is likely to output. In essence, it provides an *online* training signal that is missing from the offline pre-training set-up of modern LLMs. Our approach is also remarkably sample-efficient, yielding 38% and 64% relative increases in pass@1 rate over the zero-shot baseline and fine-tuning on MBPP data, despite fine-tuning on only 78 examples.

Our work opens up multiple avenues for promising future work. For instance, ILF can be applied iteratively over the course of multiple rounds whenever new information arrives (e.g. new Python syntax) or new bugs are discovered. As the pace of progress of modern LLM research continues to accelerate, it may soon be feasible to partially or fully automate the generation of natural language feedback (similar to ‘RL from AI feedback’ (RLAIF; Bai et al., 2022b) and our experiments in Section 3.3), greatly reducing both the time and cost necessary for collecting feedback. This direction of work is also particularly appealing because the learning signal is *process-based* rather than outcome-based, which has been shown to mitigate reward hacking and improve the correctness of intermediate reasoning steps (Uesato et al., 2022). Although further work is required to extend our method, ILF represents an exciting step forward in training LLMs with feedback that is rich, interactive, and sample-efficient.

Acknowledgements

We are grateful to Nitarshan Rajkumar, Jason Phang, Nat McAleese, Geoffrey Irving, Jeff Wu, Jan Leike, Cathy Yeh, William Saunders, Jonathan Ward, Daniel Ziegler,

Seraphina Nix, Quintin Pope, Kay Kozaronek, Peter Hase, Talia Ringer, Asa Cooper Stickland, Jacob Pfau, David Lindner, Lennart Heim, Kath Lumpante, and Pablo Morena for helpful discussions and feedback about the design and implementation of this work. We are additionally thankful to Scott Heiner and Edwin Chen for extensive help with setting up our human annotation workflow and interface. EP thanks the National Science Foundation and Open Philanthropy for fellowship support. JAC is supported by a doctoral grant from the Spanish MECD. AC, SB, and KC are supported by National Science Foundation Awards 1922658 and 2046556. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. KC is additionally supported by 42dot, Hyundai Motor Company (under the project Uncertainty in Neural Sequence Modeling) and the Samsung Advanced Institute of Technology (under the project Next Generation Deep Learning: From Pattern Recognition to AI). This project has also benefited from financial support to SB by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program), Open Philanthropy, and Apple. We also thank the NYU High-Performance Computing Center for in-kind support and OpenAI for providing access to and credits for their models via the API Academic Access Program.

References

- Austin, J., Odena, A., Nye, M. I., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C. J., Terry, M., Le, Q. V., and Sutton, C. Program synthesis with large language models, 2021. URL <https://arxiv.org/abs/2108.07732>.
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Das-Sarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T. J., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022a.
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly,

- T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional ai: Harmlessness from ai feedback, 2022b.
- Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonnell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., and Weinbach, S. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022. URL <https://arxiv.org/abs/2204.06745>.
- Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiušė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Kernion, J., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lovitt, L., Elhage, N., Schiefer, N., Joseph, N., Mercado, N., DasSarma, N., Larson, R., McCandlish, S., Kundu, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Mann, B., and Kaplan, J. Measuring progress on scalable oversight for large language models. *ArXiv*, abs/2211.03540, 2022.
- Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., and Blunsom, P. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018. URL <https://arxiv.org/pdf/1812.01193.pdf>.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code, 2021. URL <https://arxiv.org/abs/2107.03374>.
- Elgohary, A., Hosseini, S., and Hassan Awadallah, A. Speak to your parser: Interactive text-to-SQL with natural language feedback. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2065–2077, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.
187. URL <https://aclanthology.org/2020.acl-main.187>.
- Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., and Zhou, M. CodeBERT: A pre-trained model for programming and natural languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1536–1547, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.139. URL <https://aclanthology.org/2020.findings-emnlp.139>.
- Fidler, S. et al. Teaching machines to describe images with natural language feedback. *Advances in Neural Information Processing Systems*, 30, 2017.
- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- Hase, P. and Bansal, M. When can models learn from explanations? a formal framework for understanding the roles of explanation data. *arXiv preprint arXiv:2102.02201*, 2021. URL <https://arxiv.org/pdf/2102.02201.pdf>.
- Kang, S., Yoon, J., and Yoo, S. Large language models are few-shot testers: Exploring llm-based general bug reproduction, 2022. URL <https://arxiv.org/abs/2209.11515>.
- Kulal, S., Pasupat, P., Chandra, K., Lee, M., Padon, O., Aiken, A., and Liang, P. S. Spoc: Search-based pseudocode to code. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/7298332f04ac004a0ca44cc69ecf6f6b-Paper.pdf.
- Lampinen, A. K., Dasgupta, I., Chan, S. C., Matthewson, K., Tessler, M. H., Creswell, A., McClelland, J. L., Wang, J. X., and Hill, F. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*, 2022.
- Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matysińska, T., Debut, L., Bekman, S.,

- Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., and Wolf, T. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.emnlp-demo.21>.
- Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Lago, A. D., Hubert, T., Choy, P., de Masson d’Autume, C., Babuschkin, I., Chen, X., Huang, P.-S., Welbl, J., Goyal, S., Cherepanov, A., Molloy, J., Mankowitz, D. J., Robson, E. S., Kohli, P., de Freitas, N., Kavukcuoglu, K., and Vinyals, O. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022. doi: 10.1126/science.abq1158. URL <https://www.science.org/doi/abs/10.1126/science.abq1158>.
- Lin, J., Ma, Z., Gomez, R., Nakamura, K., He, B., and Li, G. A review on interactive reinforcement learning from human social feedback. *IEEE Access*, 8:120757–120765, 2020. doi: 10.1109/ACCESS.2020.3006254.
- Luketina, J., Nardelli, N., Farquhar, G., Foerster, J., Andreas, J., Grefenstette, E., Whiteson, S., and Rocktäschel, T. A survey of reinforcement learning informed by natural language. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 6309–6317. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/880. URL <https://doi.org/10.24963/ijcai.2019/880>.
- Manna, Z. and Waldinger, R. J. Toward automatic program synthesis. *Commun. ACM*, 14(3):151–165, mar 1971. ISSN 0001-0782. doi: 10.1145/362566.362568. URL <https://doi.org/10.1145/362566.362568>.
- Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., and Xiong, C. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint*, 2022.
- Odena, A., Sutton, C., Dohan, D. M., Jiang, E., Michalewski, H., Austin, J., Bosma, M. P., Nye, M., Terry, M., and Le, Q. V. Program synthesis with large language models. In *n/a*, pp. n/a, n/a, 2021. n/a.
- OpenAI. Model index for researchers, 2022. URL <https://platform.openai.com/docs/model-index-for-researchers>.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. 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=TG8KACxEON>.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems* 32, 2019.
- Pruthi, D., Bansal, R., Dhingra, B., Soares, L. B., Collins, M., Lipton, Z. C., Neubig, G., and Cohen, W. W. Evaluating Explanations: How much do explanations from the teacher aid students?, 2021.
- Scheurer, J., Campos, J. A., Chan, J. S., Chen, A., Cho, K., and Perez, E. Training language models with language feedback. *ACL Workshop on Learning with Natural Language Supervision*, 2022. URL <https://arxiv.org/abs/2204.14146>.
- Scheurer, J., Campos, J. A., Korbak, T., Chan, J. S., Chen, A., Cho, K., and Perez, E. Training language models with language feedback at scale. *Preprint*, 2023. URL https://drive.google.com/file/d/1tryv10CABT_FOF9Sn2OaWvfwqXF6iIec/view?usp=share_link.
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf>.
- Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., and Higgins, I. Solving math word problems with process- and outcome-based feedback, 2022.
- Wang, B. and Komatsuzaki, A. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- Wang, Y., Wang, W., Joty, S., and Hoi, S. C. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In *Proceedings*

of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, 2021.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. 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=_VjQlMeSB_J.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.

Workshop, B., :, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klammer, C., Leong, C., van Strien, D., Adelani, D. I., Radev, D., Ponferrada, E. G., Levkovizh, E., Kim, E., Natan, E. B., De Toni, F., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Frohberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Von Werra, L., Weber, L., Phan, L., allal, L. B., Tanguy, L., Dey, M., Muñoz, M. R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M. T.-J., Vu, M. C., Jauhar, M. A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harlman, R., Bommasani, R., López, R. L., Ribeiro, R., Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S. H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T. T., Schick, T., Thrush, T., Danchev, V., Nikoulina, V., Laippala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Taşar, D. E., Salesky, E., Mielke, S. J., Lee, W. Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler,

A., Datta, D., Szczechla, E., Chhablani, G., Wang, H., Pandey, H., Strobelt, H., Fries, J. A., Rozen, J., Gao, L., Sutawika, L., Bari, M. S., Al-shaibani, M. S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S. H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.-X., Sun, Z., Brody, S., Uri, Y., Tojarieh, H., Roberts, A., Chung, H. W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoenybi, M., Peyrounette, M., Patry, N., Tazi, N., Sansevero, O., von Platen, P., Cornette, P., Lavallée, P. F., Lacroix, R., Rajbhandari, S., Gandhi, S., Smith, S., Revena, S., Patil, S., Dettmers, T., Baruwa, A., Singh, A., Cheveleva, A., Ligozat, A.-L., Subramonian, A., Névél, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E., Winata, G. I., Schoelkopf, H., Kalo, J.-C., Novikova, J., Forde, J. Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, R., Zhang, R., Gehrmann, S., Mirkin, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Pruksachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behrooz, B., Ajibade, B., Saxena, B., Ferrandis, C. M., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D. A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J. B., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinlolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynok, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubert, S., Wang, S., Roy, S., Viguier, S., Le, T., Oyeade, T., Le, T., Yang, Y., Nguyen, Z., Kashyap, A. R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, A., Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, C., Fourier, C., Perinán, D. L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrmann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H. U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J. D., Sivaraman, K. R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M. H., Takeuchi, M., Pàmies, M., Castillo, M. A., Nezhurina, M., Sängler, M., Samwald, M., Cullan, M., Weinberg, M., De Wolf, M., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N. M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, R., Su, R., Cahyawijaya, S., Garda, S., Deshmukh, S. S., Mishra, S., Kiblawi, S., Ott, S., Sang-aaronsiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma,

T., Kusa, W., Labrak, Y., Bajaj, Y. S., Venkatraman, Y., Xu, Y., Xu, Y., Xu, Y., Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y., and Wolf, T. Bloom: A 176b-parameter open-access multilingual language model, 2022. URL <https://arxiv.org/abs/2211.05100>.

Xu, F. F., Alon, U., Neubig, G., and Hellendoorn, V. J. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, MAPS 2022, pp. 1–10, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392730. doi: 10.1145/3520312.3534862. URL <https://doi.org/10.1145/3520312.3534862>.

Zan, D., Chen, B., Yang, D., Lin, Z., Kim, M., Guan, B., Wang, Y., Chen, W., and Lou, J.-G. CERT: Continual pre-training on sketches for library-oriented code generation. In *The 2022 International Joint Conference on Artificial Intelligence*, 2022.

Ziegler, A., Kalliamvakou, E., Li, X. A., Rice, A., Rifkin, D., Simister, S., Sittampalam, G., and Aftandilian, E. Productivity assessment of neural code completion. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, MAPS 2022, pp. 21–29, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392730. doi: 10.1145/3520312.3534864. URL <https://doi.org/10.1145/3520312.3534864>.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. URL <https://arxiv.org/abs/1909.08593>.

A. Appendix

A.1. Training Details

For the experiments in Section 3.2, we run a hyperparameter sweep for all methods except for ILF. The hyperparameter value ranges that we sweep include learning rate $\in \{1.0^{-6}, 5.0^{-6}, 1.0^{-5}\}$, batch size $\in \{32, 64, 128\}$, and number of epochs $\in \{1, 2, 5\}$. The tasks for the training and validation datasets are from MBPP_{Train} and MBPP_{Refine}, respectively, while the programs are sourced from the method (*e.g.* InstructGPT, MBPP, human-written, or zero-shot CODEGEN-MONO 6.1B). For ILF, we use the best hyperparameters obtained for the sweep over MBPP programs instead of sweeping over ILF-generated programs, since the tasks in MBPP_{Refine} are already used to train π_{Refine} . All pass rates reported in Table 4 are obtained by evaluating each method on MBPP_{Test} using the best hyperparameters found during the sweep on MBPP_{Refine}.

For the experiments in Section 3.3, we separately tune hyperparameters for each size of dataset. As in our other experiments, we train and validate using the tasks from MBPP_{Train} and MBPP_{Refine}, respectively, coupled with the refinements generated by InstructGPT that pass the unit test suites. We sweep the same hyperparameter value ranges as the experiments in the previous section (*i.e.* learning rate $\in \{1.0^{-6}, 5.0^{-6}, 1.0^{-5}\}$, batch size $\in \{32, 64, 128\}$, and number of epochs $\in \{1, 2, 5\}$).

We implement all experimental pipelines with the HuggingFace transformers (v4.12.5) (Wolf et al., 2020), Huggingface datasets (v2.7.1) (Lhoest et al., 2021), and Pytorch (v1.11) (Paszke et al., 2019) libraries.

A.2. Annotator Instructions

NL Feedback for Code Generation

Given a natural language description of a Python programming challenge and its accompanying unit tests, you will be shown **10 sample model-generated Python solutions that do not pass the tests**. Please do the following:

1) Select one model-generated code sample that seems relatively easy to correct (such that it can be minimally corrected to pass the unit tests). If no such code sample exists (ie every code sample would require extensive correction, select the corresponding option and move on to the next task.

2) Write ~1-4 sentences of **natural language feedback** for the code sample that does two things: **(a) describes what is wrong with the code sample**, and **(b) how it can be fixed**. You can use individual variable or method names, but please **do not include entire lines of code**. Try to describe the necessary logic using mostly natural language, not Python expressions. Below are some examples of good versus bad feedback:

Good:

"The current code is wrong because it returns items using the heappop method in the hq module, which gets the smallest items even though the task description actually asks for the largest items. Instead, the code should use the nlargest function in hq to get the largest n integers."

Bad (because it only describes what to change, and not what was originally wrong with the code):

"The code should use the nlargest function in hq to get the largest n integers."

Bad (because it gives the actual code needed to fix the function, instead of describing the desired fix using words only):

"The current code is wrong because it returns items using the heappop method in the hq module, which gets the smallest items even though the task description actually asks for the largest items. Instead, the function should return [hq.heappop(nums) for i in range(n)]."

3) Write a **refinement** for the code that consists of copying the original model sample and making **minimal edits** to the code such that it correctly solves the challenge and can pass all the unit tests.

Here is an example. We only show a couple of model samples here for brevity, but in practice you will be shown up to 10 model samples per task.

Figure 6. An example of the instructions given to Surge annotators.

A.3. Examples of Human Versus InstructGPT Feedback

Table 7: Examples of natural language feedback written by humans versus InstructGPT for incorrect programs generated by CODEGEN-MONO 61.B.

Incorrect program generated by CODEGEN-MONO 61.B	Human feedback	InstructGPT feedback
<pre> """ Write a function to assign frequency to each tuple in the given tuple list. >>> Example: assign_freq([(6, 5, 8), (2, 7), (6, 5, 8), (6, 5, 8), (9,), (2, 7)]) = '[(6, 5, 8, 3), (2, 7, 2), (9, 1)]' """ from collections import Counter def assign_freq(test_list): """ >>> assign_freq([(6, 5, 8), (2, 7), (6, 5, 8), (6, 5, 8), (9,), (2, 7)]) = '[(6, 5, 8, 3), (2, 7, 2), (9, 1)]' """ counter = Counter(x for x in test_list) res = [] for key, val in counter.items(): temp = (key, val) res.append(temp) return str(res) </pre>	<p>“The only issue with this code is that it does not dereference the key. When you set temp = (key, val), the key is an entire tuple instead of the values in the tuple. Thus, you need the asterisks to get the values”</p>	<p>“The function works as intended.”</p>
<pre> """ Write a function to get the angle of a complex number. >>> Example: angle_complex(0,1 j)=1.5707963267948966 """ import cmath def angle_complex(a,b): return cmath.phase(complex(a,b)) </pre>	<p>“Only one thing is wrong in the code that complex method can take the complex number as argument example complex(a+b) not like complex(a,b) we have to replace, by + and the code is fine.”</p>	<p>“The code should use the atan2 function to get the angle of a complex number.”</p>

Table 7: Continued from previous page

Incorrect program generated by CODEGEN-MONO 6.1B	Human feedback	InstructGPT feedback
<pre> """ Write a function to count the same pair in three given lists. >>> Example: count_samepair ([1,2,3,4,5,6,7,8], [2,2,3,1,2,6,7,9], [2,1,3,1,2,6,7,9])=3 """ def count_samepair(list1,list2 ,list3): count = 0 for i in range(len(list1)) : for j in range(len(list2)): if list1[i]==list2 [j] and list2[j]==list3[i]: count+=1 return count </pre>	<p>“One extra for loop for list2 is causing the issue, it finds more matching elements between list1,2,3. The possible solution for the code is to remove the unnecessary for loop of list2 iteration, and check all the lists same index element at the time, and return the count of this element.”</p>	<p>“The code above is very inefficient. It is checking every possible combination of pairs of elements in the three lists. It would be much more efficient to use a set to keep track of the pairs that have already been seen.”</p>
<pre> """ Write a function to find number of lists present in the given tuple. >>> Example: find_lists([1, 2, 3, 4], [5, 6, 7, 8])) = 2 """ def find_lists(Input): result = 0 for x in Input: if isinstance(x, list) : result += 1 return result </pre>	<p>“The code works for tuples of more than one list, but not in case of one single list. Python goes one level down and reads the elements when given a single list and does not see an instance of list. I would just check for the case of ‘type(Input) == list’ and turn the input into a 1-sized list of lists in this case.”</p>	<p>“The function find_lists needs to return the number of lists present in the given tuple, not the number of items in the tuple.”</p>

Table 7: Continued from previous page

Incorrect program generated by CODEGEN-MONO 6.1B	Human feedback	InstructGPT feedback
<pre> """ Write a function to calculate magic square. >>> Example: magic_square_test ([[7, 12, 1, 14], [2, 13, 8, 11], [16, 3, 10, 5], [9, 6, 15, 4]])=True """ def magic_square_test(my_matrix): """ >>> magic_square_test([[7, 12, 1, 14], [2, 13, 8, 11], [16, 3, 10, 5], [9, 6, 15, 4]])=True """ def is_square(x): return x==x*x def is_magic_square(my_matrix): sides=[] for i in range(len(my_matrix)): sides.append(sum(my_matrix[i])) if not is_square(sum(sides)): return False sides=[] for i in range(len(my_matrix)): sides.append(sum([my_matrix[i][j] for j in range(len(my_matrix))])) if not is_square(sum(sides)): return False diagonals=[] for i in range(len(my_matrix)): diagonals.append(my_matrix[i][i]) if not is_square(sum(diagonals)): return False diagonals=[] for i in range(len(my_matrix)): diagonals.append(my_matrix[i][len(my_matrix)-1-i]) if not is_square(sum(diagonals)): return False return True return is_magic_square(my_matrix) </pre>	<p>“The logic behind checking whether the sum of rows, columns, and diagonals is a magic number or not is not correct, the code checks if the sum of squares of itself in the is_square method is wrong. Instead of checking all the sums of rows, columns, and diagonals at every point, the logic should be like saving all the sums for rows, columns, and diagonals in one array(sides) and then checking if there are all same values present in this array because for magic_square_test all the sums should be equal.”</p>	<p>“The function magic_square_test should return whether my_matrix is a magic square.”</p>

Table 7: Continued from previous page

Incorrect program generated by CODEGEN-MONO 6.1B	Human feedback	InstructGPT feedback
<pre> """ Write a python function to count characters at same position in a given string (lower and uppercase characters) as in english alphabet. >>> Example: count_char_position(" xbcefg") = 2 """ def count_char_position(str1): counter = 0 for i in range(0, len(str1)): if str1[i] == str1[i]. upper(): counter += 1 return counter </pre>	<p>“a) The first thing that the code sample gets incorrect is that it increments the counter when each character in the string is an upper case character. The prompt is asking for how many characters in str1 have the same numerical position as their respective alphabetic counterpart and the code does not check for this at all. b) The way to fix it would be to create a separate j variable that keeps track of the ordinal value that corresponds with the alphabetic character that we would expect to be compared to the respective character in str1. j and i would increment at the same time so that the code is both iterating through the alphabet AND iterating through str1 simultaneously. The other fix would be to automatically convert str1 to lowercase so that we don’t have to worry about checking the ordinal codes of both the lowercase and uppercase versions of the same alphabetic character.”</p>	<p>“Great job!”</p>