

# DeepPERF: A Deep Learning-Based Approach For Improving Software Performance

Spandan Garg\*  
spgarg@microsoft.com  
Microsoft  
Redmond, Washington, USA

Roshanak Zilouchian  
rozilouc@microsoft.com  
Microsoft  
Redmond, Washington, USA

Colin B. Clement  
coclemen@microsoft.com  
Microsoft  
Redmond, Washington, USA

Neel Sundaresan  
neels@microsoft.com  
Microsoft  
Redmond, Washington, USA

Chen Wu†  
overwindows@hotmail.com  
Microsoft  
Shanghai, China

## ABSTRACT

Improving software performance is an important yet challenging part of the software development cycle. Today, the majority of performance inefficiencies are identified and patched by performance experts. Recent advancements in deep learning approaches and the wide-spread availability of open source data creates a great opportunity to automate the identification and patching of performance problems. In this paper, we present DeepPERF, a transformer-based approach to suggest performance improvements for C# applications. We pretrain DeepPERF on English and Source code corpora and followed by finetuning for the task of generating performance improvement patches for C# applications. Our evaluation shows that our model can generate the same performance improvement suggestion as the developer fix in ~53% of the cases, getting ~34% of them verbatim in our expert-verified dataset of performance changes made by C# developers. Additionally, we evaluate DeepPERF on 50 open source C# repositories on GitHub using both benchmark and unit tests and find that our model is able to suggest valid performance improvements that can improve both CPU usage and Memory allocations. So far we’ve submitted 19 pull-requests with 28 different performance optimizations and 9 of these PRs have been approved by the project owners.

## 1 INTRODUCTION

Performance bugs are usually non-functional bugs that can cause poor user experience, reduced throughput, increased latency, and wasted resources. Performance bugs may not cause system failure and may depend on user input, therefore detecting them can be challenging [8, 13]. They also tend to be harder to fix than non-performance bugs [17, 24]. As a result, better tool support is needed for fixing performance bugs.

In recent years, a variety of performance bug detection approaches have emerged to help developers identify performance issues. However, a majority of existing performance bug detection approaches focus on specific types of performance problems, often relying on static code analysis or some pre-defined set of rules to detect performance issues. For instance, prior work investigated the detection of inefficient loops [23, 17, 28], database related performance issues, low-utility data structures [30], false sharing specially in

multi-threaded code [15], etc. Building a rule-based analyzer is a non-trivial task as it requires achieving the right balance between precision and recall. Once developed, maintaining these rules can also be costly [2] as it requires continuous effort by performance experts.

With the recent rise of large transformer models and wide-spread availability of open-source software artifacts, there is an opportunity to learn patterns of performance improvements directly from mined data. Transformer-based approaches have been shown to achieve state-of-the-art performance, not only in various Natural Language Processing (NLP) problems, but also a variety of software engineering tasks such as code-completion [25], documentation generation [5], unit test generation [26], bug detection [9], etc. In this paper, we draw inspiration from these techniques, in an attempt to solve the problem of automatically suggesting performance improvements.

We present an approach called DeepDev-Perf that uses a large transformer model to suggest changes at application source code level to improve its performance. We first pretrain our model using masked language modelling (MLM) tasks [5] on English text and source code taken from open source repositories on GitHub, followed by finetuning on millions of performance commits made by .NET developers. Through our evaluation, we show that our approach is able to recommend patches to provide a wide-range of performance optimizations in C# applications. Further, by suggesting changes to a set of real world repositories and measuring the impact of our suggestions through benchmark tests, we show that our changes provide actual performance gains to these applications. 9 PRs containing our model suggestions have already been accepted by the developers of these projects, showing that our suggestions are considered to be correct and useful by the project owners.

In summary, our work makes the following main contributions:

- We propose a novel transformer-based model called DeepPERF, which finds performance optimization opportunities in a c# application and automatically generates performance improvement patches.
- We extensively evaluate DeepPERF using a curated dataset of real-world performance improvement changes made by C# developers to a hold-out set of open source repos on GitHub. Through our empirical evaluation, we demonstrate that Deepdev-PERF is able to generate a wide-variety of performance improvements.

\*Corresponding Author

†Work done at Microsoft.

- We show real-world evidence that DeepPERF generates changes that lead to tangible performance improvements to various open source C# projects on GitHub. We submit PRs containing the suggested changes to these repos, many of which have since been approved showing that our fixes are considered useful by developers.

## 2 MOTIVATING EXAMPLES

Figure 1 shows examples of two suggestions made by DeepPERF to two open-source C# projects on GitHub. In the first example, the code prior to the change uses LINQ [19]. LINQ expressions have an inherent allocation associated with them. As a result, LINQ usage on the application hot-path often leads to unnecessary allocations, which can cause spikes in garbage collection (GC), depriving the application of CPU resources and reducing its throughput. In the top change, DeepPERF recognizes that the use of LINQ call to skip the first position (Line 463) is unnecessary and it recommends a change to unroll the LINQ query and use an explicit for-loop, which starts indexing from 1. By executing the unit and benchmark tests in this repository, we verified the correctness of the change as well as the performance gain. Looking at the benchmark results, the change achieves a reduction in allocations as well as Gen 0 GC compared to prior code.

The second change shows a repeated array allocation happening inside a while-loop, within a recursive method. DeepPERF notices that the array being allocated is independent of the loop and method call and suggests a change to hoist the array as a private static member of the containing class. This ensures that only one instance of the array is ever allocated, which reduces the application’s allocations.

Pull-requests containing both of these changes were submitted to the corresponding GitHub repos and have since been approved by their owners.

## 3 OUR APPROACH

We describe the details of our proposed model below. Figure 2 shows an overview of our pipeline. We begin by first describing how we take an English-pretrained BART-large model and further pretrain it on Source code. We then describe our data collection and example generation pipeline for finetuning. This is followed by a description of our two-step finetuning process using the examples generated by the example generation step.

### 3.1 Pretraining

For pretraining, we collected 45K GitHub repositories with  $\geq 5$  stars that were composed primarily of C# code. We de-duplicate this data on a file-level using a hash function and then pretrain a 406M parameter BART-large model using span-masking objective [14] on this data set. Span-masking corrupts code by replacing random spans of input tokens with a <MASK> token, and the model is trained to reconstruct the original code by predicting the tokens replaced by the mask tokens. Such pretraining has been shown to significantly improve model performance for a various downstream NLP tasks, including many software engineering tasks [5].

<sup>1</sup><https://github.com/CreoOne/V/pull/1>

<sup>2</sup><https://github.com/tihlv/Omtt/pull/1>

The figure displays two code snippets with suggested changes highlighted in red and green. The top snippet is from `V/V/Vector.cs` and shows a change from a LINQ query to an explicit for-loop. The bottom snippet is from `Omtt/Omtt.Parser/Lexical/TemplateModelLexicalParser.cs` and shows a change to hoist a static array allocation.

```

V/V/Vector.cs
456 private void EnsureConsistentDimensionality(Vector[] vectors)
457 {
    // ...
463     foreach (Vector vector in vectors.Skip(1))
464     if (dimensions != vector.Dimensions)
465         throw new DimensionalityMismatchException();
466 }

456 private void EnsureConsistentDimensionality(Vector[] vectors)
457 {
    // ...
463     for (int i = 1; i < vectors.Length; i++)
464     if (dimensions != vectors[i].Dimensions)
465         throw new DimensionalityMismatchException();
466 }

Omtt/Omtt.Parser/Lexical/TemplateModelLexicalParser.cs
45 private void ProcessContent(ParsingSource source, List<Lexem>
    result, bool textMode, String[] exitSymbols)
46 {
    var symbols = (textMode ? TextLiterals : SymbolLiterals)
        .Union(exitSymbols)
        .ToArray();

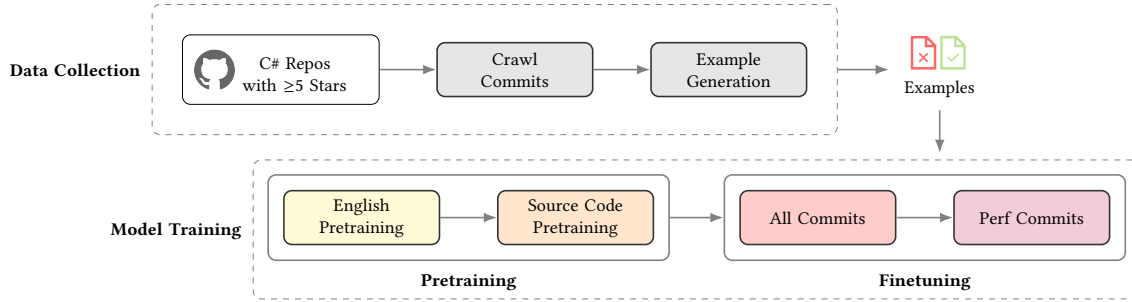
49
50 Boolean foundExitSymbol = false;
51 while (source.Any() && !foundExitSymbol)
52 {
    // ...
67     else if (!foundExitSymbol && symbol !=
        MarkupLiterals.CloseSymbol && textMode)
68     ProcessContent(source, result, false, new []
        {MarkupLiterals.CloseTagSymbol,
        MarkupLiterals.CloseExpression});
69 }
70 }

25 private static readonly String[] CloseTagLiterals = new[]
26 {
27     MarkupLiterals.CloseTagSymbol,
28     MarkupLiterals.CloseExpression
29 };
    // ...
51 private void ProcessContent(ParsingSource source, List<Lexem>
    result, bool textMode, String[] exitSymbols)
52 {
    var symbols = (textMode ? TextLiterals : SymbolLiterals)
        .Union(exitSymbols)
        .ToArray();

54
55 Boolean foundExitSymbol = false;
56 while (source.Any() && !foundExitSymbol)
57 {
    // ...
73     else if (!foundExitSymbol && symbol !=
        MarkupLiterals.CloseSymbol && textMode)
74     ProcessContent(source, result, false,
        CloseTagLiterals);
75 }
76 }

```

**Figure 1: Two examples of the kinds of changes DeepPERF is able to suggest: (i) the change on top, taken from a PR<sup>1</sup> submitted using our DeepPERF model to a C# project on GitHub. The change is to unroll a LINQ query into an explicit for-loop. This change results in lower allocations and Gen 0 garbage collection compared to prior code. (ii) the change on the bottom is from another PR<sup>2</sup> to a different C# project. In this change, the code is repeatedly allocating the same array inside a loop. Since the array is independent of the loop or the method itself, DeepPERF suggests a change to hoist the array to a static member of the containing class, which results in a reduction in allocations.**



**Figure 2: Our model data collection and training pipeline.** We first crawl the commit history of all 45k C# repos with  $\geq 5$  stars on GitHub and generate examples for each modified method with various contextual elements important to performance (using statements, class attributes, caller-callee methods, etc.). For training, we first pretrain BART-large on denoising objectives over English text and source code, followed by a two-step finetuning. In our finetuning step, we first finetune the model on examples generated from all commits, followed by a smaller finetuning step done only on examples from commits where developer included a performance related keyword ("perf", "performance", "reduce allocation", etc.) in the commit message.

### 3.2 Data Collection

Below we present the details of our data collection and example generation steps. For our data, we collect  $\sim 45k$  repositories with  $\geq 5$  stars on Github whose primary language is C# and had a commit within the last 5 years.

**3.2.1 Crawling Commits.** After cloning these projects, we crawl the *main* branch’s commit history. This history involves a commit message and a diff representing the difference between current and previous version of the changed files. This yields  $\sim 11M$  commits and their corresponding commit messages.

**3.2.2 Identifying Performance-related Commits.** For each commit, we look for performance related keywords ("perf", "performance", "reduce allocation", etc.) in the commit message to determine if it is performance related. Table 1 shows the number of commits and examples that come from performance-related commits.

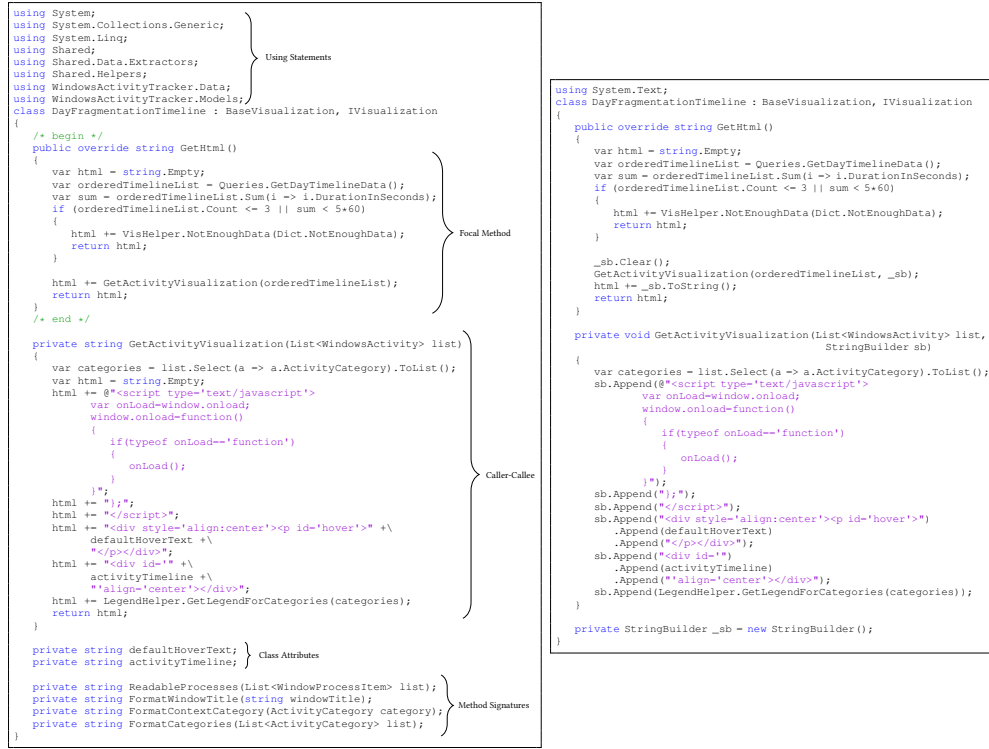
**3.2.3 Generating Code Transformation Pairs.** To generate the code transformation pairs (input/output sequences) from the crawled commits, we follow the following process. Within each commit, we parse the modified C# files that end with the extension ".cs" using the tree-sitter parser. We first extract the classes within the file and its corresponding member methods. We apply some pre-processing steps on the method bodies by normalizing white-space and removing comments. This allows us to ignore any trivial modifications to whitespace and comments. Using the method signature, we identify the corresponding versions of the method in the before and after files. We then compare the normalized method bodies between the two versions of the file and discards the methods whose bodies do not appear to be have been modified. From here on, we refer to the remaining modified methods as focal methods.

Next we construct an input/output pair for each focal method. Figure 3 shows an example of one such input/output pair. We start by including the focal method itself in the example input, whose location is indicated to the model using C-style comments (`/* edit */` and `/* end */`) before and after the focal method. Performance changes often require changes beyond the focal method itself (as seen in the second example in Figure 1), such as adding

new class level attributes or additional imports, or even changes to other methods within the class that make calls to or are called by the focal method. Prior work has shown that including additional class/file-level context information with the focal method results in a higher quality predictions in such code generation tasks [26, 9]. We believe that such contextual information would prove useful in generating performance patches as well. Below are the file/class level context elements we include as part of the input:

- *Using Statements:* These tell the model what import statements exist within the file and whether new imports need to be added for the new methods or APIs used in the recommended changes.
- *Class Attributes:* These are the containing class’s member attributes. The underlying types of class attributes can often be important information in determining the right performance fix. This may include fixing incorrect usages of variables of certain types that cause performance issues or recommending a more appropriate data structure e.g. replacing `List<T>` with a `HashSet<T>`, etc.
- *Caller-Callee Methods:* These are the methods that directly make calls to or are, in turn, called by the focal method. This information can help the model learn patterns of changes that involve hoisting/memoizing calls across methods, to optimize computations or simply modifying the caller/callee to be consistent with the modified focal method.
- *Method Signatures:* Finally, we add the signatures for any other methods that aren’t caller or callee methods of the focal method. Due to limited token space, we are unable to add the bodies of each method. This information could shed some light on the nature of the class itself and provide context as to what other methods are present in the class for the model to use in the generated patch.

Due to the input token window for BART-large being limited to only 1024 tokens, we construct the example input in an iterative fashion. We start by including the focal method in the example input and then incrementally add each contextual element in the list above. Before adding each type of contextual elements, we



**Figure 3: An example of an Input-output pair used in finetuning.** The input consists of the focal method along with various class/file level context elements such as using statements within the file, class member attributes, focal method’s caller/callee methods, other method signatures. We also include C-style comments (*/\* begin \*/* and */\* end \*/*) before and after the focal method, indicating its location to the model. The example shown above comes from an actual performance commit made by a C# developer to an open source repo and shows a perf change to replace a sequence of string concatenations with a `StringBuilder`. Such a change would save allocations as each string concatenation leads to a new string allocation, whereas the `StringBuilder` defers the allocation until after all the component strings are gathered and a call to `ToString()` is made. Additionally, the change also caches and re-uses the `StringBuilder` instance, as opposed to allocating a new one each time. In this case, the output also consists of an additional using statement, importing the namespace containing the definition of `StringBuilder`, along with modified versions of the focal and callee method and the new class attribute.

ensure that the resulting sequence will be within the allowed range of tokens i.e.  $\leq 1024$  tokens.

For the output, in addition to changes to the focal method, we include any of focal method’s caller-callee methods that are modified by the commit. We also include any additional imports that may have been added as well as class attributes defined/modified that are used by the focal method or modified caller-callee methods. This way we allow the model to output patches that make changes to not only the focal method but also the caller-callee methods, class attributes as well as add any new methods, attributes and import statements as needed. Figure 3, shows an example of an input output pair generated using the steps above.

**3.2.4 Data splitting and De-duping.** We split the finetuning data on the project level. We leave out two sets of test and validation repos, each containing 600 repos that are not included in either step’s training data. We also dedupe the examples in each set as well as remove any near duplicates [1] among them to ensure no overlap between train and test data.

**Table 1: Number of commits and examples in our training data for the All Commits and Perf Commits finetuning steps.**

Commit Type	Data	# of Commits	# of Examples
All Commits		11M	16M
Perf Commits		535k	1.5M

### 3.3 Finetuning

We finetune the code pretrained BART-large model on the task of generating a performance improvement code transformation, given an input sequence containing the focal method and contextual elements (as explained in Section 3.2.3). We perform a two-stage finetuning. We first finetune our pretrained transformer model on examples from all commits. We refer to the resulting model as DeepDev-C#. We then perform a second finetune step over DeepDev-C# model using the set of code transformations examples extracted from performance commits, to teach it specifically how to make performance changes. We refer to the final model as

DeepPERF. To better understand whether the first finetuning step has an impact on the results, we train a third model, finetuned only on code transformations extracted from performance commits. We refer to this model as DeepDev-Perf-Commits. In our evaluation, we compare these three models above and discuss possible reasons for differences in their performances.

## 4 EMPIRICAL EVALUATION

In this section, we first explain our baselines and evaluation metrics. We then cover our quantitative and qualitative analysis methodology and results.

### 4.1 Baselines and Evaluation Metrics

We compare DeepPERF model with the following two models:

- **DeepDev-C#**, which was first pretrained on English and Source code, followed by finetuning on code transformations extracted from all C# commits in our training data.
- **DeepDev-Perf-Commits**, which was first pretrained on English and Source code, followed by finetuning on code transformations extracted from only performance commits (commits with a performance-related keyword in the commit message) in our training data.

In order to compare the models' performances we define the following metrics:

- **Verbatim Match %**: We report the number of examples where one of the model's suggestions was found to match verbatim with the developer patch i.e. the ground truth output in the input/output pair.
- **Abstracted Match %**: For our comparison to be independent of variable name matching, we replace variable names with generic names of the form VAR\_*i* (e.g. VAR\_0, VAR\_1, etc.), where "i" is determined based on the relative order in which variables are encountered when traversing the parse tree. We then compare the abstracted versions of the model suggestions with the similarly abstracted developer patch and report how many were found to match.
- **CodeBLEU**: We measure the CodeBLEU [20] scores to gauge the similarity of model output with the actual developer patches. In addition to n-gram matching of BLEU, CodeBLEU also compares abstract syntax trees (AST) and data flow between two programs. Thus, it takes into account the syntactic as well as semantic code similarity. We use the hyperparameters that were shown to have the highest correlation to human scores in the study i.e.  $\alpha, \beta, \gamma, \delta = 0.1, 0.1, 0.4, 0.4$ .

Through this experiment, we intend to answer the following research questions:

- **RQ1**: Are both finetuning steps (All Commits and Perf Commits step) in our two-step finetuning necessary?
- **RQ2**: Is DeepPERF able to provide a wide-range of performance optimizations?

### 4.2 Quantitative Analysis

For the purposes of our evaluation, we picked a random set of repos from our test set, none of which had previously been seen by

our models. We then collected all performance commits (commits with a performance-related keyword in the commit message) that change only one ".cs" file to filter out squash merges that include a variety of changes across multiple files. This process yields ~1500 commits, resulting in a total of ~2100 examples. We use this set of examples for our evaluation.

**4.2.1 Two-step Finetuning Ablation.** Table 3 shows the results of our 3 models over this dataset using the metrics we defined earlier. We see that our DeepPERF model performs better than the other two models and is able to get ~16% of the examples in this dataset verbatim and ~19% verbatim when the variable names are abstracted away. The 3 models achieve very similar CodeBLEU scores, which is expected since, due to the nature of the task, most of the code will be shared between the model output and ground truth for any well-trained model. The major difference between DeepPERF and the DeepDev-Perf-Commits is that DeepPERF is first finetuned on examples generated from all C# commits, followed by a smaller finetuning step on examples generated using commits with a performance related keyword in commit title/description. On the other hand, DeepDev-Perf-Commits is finetuned directly on the smaller set of performance commit examples. DeepDev-C# differs from the two as it's only finetuned on all C# commits, but not directly on performance commits.

Comparing the results of the three, the reason both DeepPERF and DeepDev-C# perform better than DeepDev-Perf-Commits could be because finetuning on all commits allows the models to learn better representations for code by seeing more examples of how changes are made by C# developers, since the dataset for All Commits step is almost 10 times larger than the one containing only examples from perf commits. There is also a possibility that there may be some performance improvement changes that aren't explicitly annotated within the commit message. For example, developers may not always explain every change they make in their commits and squash multiple changes into a single commit, mentioning only the most important in the commit message. We found several examples of such commits in our training data, where the changes contained performance optimizations, but the commit message did not include any performance related discussion. While, we don't know pervasive this is and just how many such "phantom" performance changes exist outside of performance commits, their presence would imply that models trained on All Commit data would overall be seeing more performance improvement code transformations during training than one that's been directly trained on performance commits.

### 4.3 Qualitative Analysis

To better understand the different types of performance improvements DeepPERF can suggest, we performed a qualitative evaluation on a subset containing ~125 performance commits from our test set in the previous section. This resulted in a set of 132 examples demonstrating a variety of performance fixes, each of which were verified with two C# performance experts. Based on our understanding of performance changes in C#, we observe that the changes fall into the following three broad categories of performance improvements:

**Table 2: Three categories of performance issues in expert-verified dataset of performance improvements.**

Change Category	Examples of Performance Optimizations Within Category	# of Examples
High Level Change (C1)	Memoize results using <code>Dictionary/ConcurrentDictionary</code> , Hoist computation/allocation to outer scope (loop, method, class, etc.), Introduce fast-path to avoid unnecessary computation, Re-use types like <code>List</code> , <code>Dictionary</code> , <code>StringBuilder</code> , etc.	29
Suggest Different API/Data Structure (C2)	Replace a series of string concatenations with a <code>StringBuilder</code> , <code>List&lt;T&gt;.Contains() → HashSet&lt;T&gt;.Contains()</code> , Remove LINQ usage (e.g. <code>List&lt;T&gt;.Any() / Count() → List&lt;T&gt;.Count</code> , etc.), etc.	71
Improve Existing API/Data Structure Usage (C3)	Condense/Optimize LINQ queries (e.g. <code>Count() == 0 → !Any()</code> , <code>Where(&lt;lambda&gt;).FirstOrDefault() → FirstOrDefault(&lt;lambda&gt;)</code> , etc.), <code>Dictionary&lt;K, V&gt;.ContainsKey() → Dictionary&lt;K, V&gt;.TryGetValue()</code> , <code>return new List&lt;T&gt;() → return Array.Empty&lt;T&gt;()</code> , <code>string.Equals() → string.Length == 0</code> , etc.	34

**Table 3: Summary of the results of our three models over dataset comprising of all the examples generated from commits that had a performance related keyword in the commit message, in a hold-out set of 50 test repos.**

Model	Verbatim Match %	Abstracted Match %	CodeBLEU
DeepDev-C#	14.2	16.5	69.2
DeepDev-Perf-Commits	13.3	16.4	68.8
DeepPERF	15.6	19.1	69.3

- (Category 1) High Level Changes: These consist of algorithmic changes that require modifications to the overall code structure. These changes could include hoisting calls/allocations to an outer scope, adding caching/memoization to avoid repeated computation, introducing a fast-path, etc.
- (Category 2) Suggesting Different API/Data Structure: These are language/API specific changes to replace or remove an existing API or data structure usage in favor of a better alternative. These changes could include removing LINQ by unrolling queries into explicit loops, suggesting a different data structure (e.g. replace `List` with a `HashSet`, when performing look-ups), etc.
- (Category 3) Improving Existing API/Data Structure Usage: These are also language/API specific changes, but suggest modifications to existing usage of an API or data structure when deemed incorrect or sub-optimal. These may include changes like condensing LINQ queries to be more optimal, fixing incorrect uses of a data structure, using a better suited overload of a library function, etc.

Table 2 shows some example performance changes found within the 3 categories as well as the number of examples in our dataset that fell within that category. We plan to release this dataset with each example labelled with above categories as an artifact upon acceptance of this work.

**4.3.1 Human Evaluation Methodology.** For each example in our dataset, we sample 2000 hypotheses from the model and take the top 500 suggestions, based on the average likelihood of tokens. Since we have so many suggestions, 500 suggestions for >100 examples, it would be difficult to evaluate each of them by hand even with a team of experts. Therefore, we use the following evaluation metric to help us approximate the model’s Top-K Accuracy. We report this in addition to the metrics mentioned earlier (i.e. Verbatim Match, Abstracted Match and CodeBLEU):

- **Closest Match Top-K Accuracy %:** Using a code search technique such as Aroma [16], we find the document within the corpus of model suggestions that is closest to the developer patch, for each example. We then verify the most similar suggestion with two performance experts, neither of whom are on the author list. The experts are shown both the developer change and model suggestion and asked to assess whether they considered the model suggestion to be making the same performance optimization and are semantically the same as the developer change.

**Table 4: Summary of the results of our three models over the manually curated dataset.**

Model	Closest Match Top-K Accuracy %				Verbatim Match %	Abstracted Match %	CodeBLEU
	1	10	100	500			
DeepDev-C#	2.3	14.4	31.1	37.9	24.6	26.1	68.3
DeepDev-Perf-Commits	7.6	18.9	31.2	42.4	26.1	29.1	70.6
DeepPERF	8.3	18.2	34.1	53.0	34.3	37.3	70.7

**4.3.2 Results.** Table 4 and Figure 4, show the results of the 3 models using our defined metrics. We see that our best model is able to solve ~53% of the examples in our dataset, getting ~34% verbatim as the developer fix. The Closest Match Top-K Accuracy plot was computed based on the associated rank of the retrieved suggestion among the top-500 hypotheses, when found to be correct by both of the two C# performance experts. In majority of the cases, the



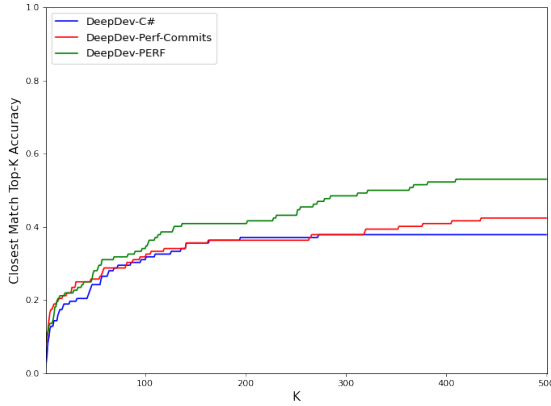


Figure 4: Closest Match Top-K Accuracy plot of our models on the manually curated dataset of performance changes. We can see that DeepPERF achieves the best score among the 3 models.

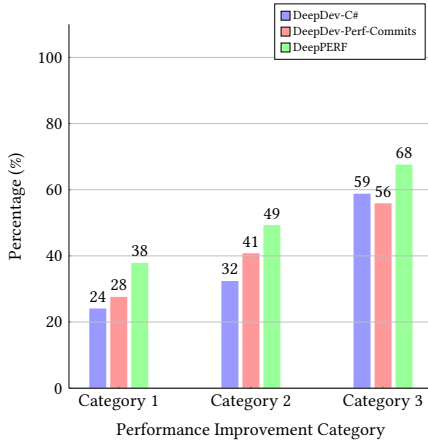


Figure 5: Performance of our models on the three categories of performance issues: High Level Changes (Category 1), Suggesting Different API/Data Structure (Category 2), Improving Existing API/Data Structure Usage (Category 3). We can see that the DeepPERF model (green) tends to perform the best among the models in all three categories, followed by DeepDev-Perf-Commits and DeepDev-C#.

main reasons for dissimilarities from the developer were the model suggesting different variable names, or other slight variations like using the `var` keyword instead of the variable's type or using a `for`-loop as opposed to a `foreach`-loop where both are appropriate, difference in order of statements where relative order did not matter (such as using statements at the start of file), etc. Figure 5 shows the performance of our models in the 3 categories of performance

changes. We also see that our best model out-performs the other two models in each category.

**RQ1 & RQ2:** In summary, we conclude that both finetuning steps were necessary as DeepPERF clearly demonstrates better performance than the other two models on the overall test dataset of performance commits as well as the smaller expert verified dataset of performance optimizations. Additionally, we see that DeepPERF can generate suggestions that span a wide variety of performance optimizations encompassing both high-level algorithmic to low-level API/Data structure related performance changes. Furthermore, these changes were considered equivalent to the developer-made performance improvements by two performance experts in C#.

```
// Summary
BenchmarkDotNet v0.13.1, OS=Windows 10.0.19044.1526 (21H2)
Intel Core i7-6600U CPU 2.60GHz (SandyLake), 1 CPU, 8 logical and 2 physical cores
.NET SDK 6.0.102
DefaultJob : .NET 6.0.2 (6.0.222.6466), X64 RyuJIT
DefaultJob : .NET 6.0.2 (6.0.222.6466), X64 RyuJIT
```

Method	Mean	StdDev	StdErr	Median	Q1	Q3	Iterations	Gen 0	Allocated
SpanHeadStart	0.155 ns	0.111 ns	0.038 ns	0.055 ns	0.000 ns	0.086 ns	13,000	-	-
SpanHeadLast	2,210,277 ns	13,7827 ns	9,3472 ns	2,288,403 ns	2,187,907 ns	2,232,437 ns	13,000	-	-
SpanHeadFirst	4,426,247 ns	27,3170 ns	18,3240 ns	4,426,402 ns	4,280,112 ns	4,576,500 ns	14,000	12,0032	25,4688 B
SpanHeadEnd	13,524,802 ns	180,9718 ns	28,8787 ns	13,524,185 ns	13,402,039 ns	13,597,986 ns	14,000	25,4340	54,792 B
SpanHeadLength	14,402,108 ns	2,497,4620 ns	325,4322 ns	14,402,722 ns	13,986,706 ns	15,046,412 ns	48,000	28,4308	54,792 B
BinaryHeaderReadPerformance	148,984,707 ns	26,424,4780 ns	2,713,1330 ns	149,977,851 ns	128,066,223 ns	159,924,929 ns	95,000	60,1051	261,240 B

Figure 6: An example of the output generated from executing a BenchmarkDotNet test suite. The first column shows the different benchmarks defined by the user. BenchmarkDotNet automatically runs each of these user-defined benchmarks multiple times and reports metrics such as sample mean, standard deviation, standard error, median, first and third quartile of the test duration. For allocations, it reports the memory allocated on average during each test run.

## 5 IN-THE-WILD EVALUATION

To see whether our approach can suggest real performance improvements to existing C# projects, we performed an "in-the-wild" evaluation of DeepPERF on a set of 50 C# GitHub repos, not previously seen by our model. For this evaluation, we chose repos that contained both benchmark and unit tests to allow us to verify that our suggested changes are correct (using unit tests) and lead to measurable performance improvements (using benchmark tests). Benchmark tests in C# are usually written using the BenchmarkDotNet library. Shown in Figure 6 is an output summary of a BenchmarkDotNet test. BenchmarkDotNet automatically runs each benchmark test in the user-written test suite multiple times and reports metrics for the duration of a given benchmark test as well as the amount of memory that is allocated on average on each run. It can also give other information such as how frequently Generational GC is triggered.

Through this experiment, we intend to address the following research questions:

- **RQ3:** What are the reasons behind some of our model's changes failing to compile? How could this be improved?
- **RQ4:** Is DeepPERF able to suggest changes that lead to real performance improvements? If so, how much performance improvement do these changes typically provide? Are these suggestions considered useful by the developers?

- **RQ5:** How effective are unit and benchmark tests in ensuring changes are correct performance improvements?

## 5.1 Experiment Setup

To find repos with Benchmark tests, we sampled 50 C# repos where BenchmarkDotNet NuGet package is mentioned in one of the build configuration files ("\*.csproj") in the repo. To limit the methods we need to generate changes for, we select methods that satisfy following two criteria: (1) are present on the execution path of the repo's benchmark tests, (2) have a high line/branch coverage with the repo's unit tests. This yields 201 methods across the 50 test repos. We then generate model inputs for each of these methods, including contextual information as described in Example generation step (Sec. 3.2.3). We use our best model, DeepPERF, and sample 2000 suggestions for each of these inputs. We then pick the top 100 suggestions based on their average token likelihood, and test each of these suggestions against the repo's main branch.

From these suggestions, we first filter out changes that are syntactically incorrect. We found that ~7% of the suggestions had a syntax error. Most of these were due to early truncation or repetition when generating long outputs, which are known issues when generating text using such language models.

## 5.2 Running Unit Tests

We then run unit tests for each of the remaining ~93% suggestions. This step filters out suggestions that fail to compile or are found incorrect based on the unit test cases provided by the developer. Table 5 shows a breakdown of how many suggestions fail at this stage. As we can see, at the end of this step we are left with ~44% of the suggestions we started with.

**Table 5: Breakdown of the results of running unit tests.**

Result	Occurrences	% of Suggestions
Syntax Error	1329	6.6
Compilation Error	7860	39.1
Failed Unit Tests	2056	10.2
Passed Unit Tests	8855	44.1
Total	20100 = 201 * 100	100%

## 5.3 Analyzing Build Failures (RQ3)

Table 6 shows the main reasons of compilation errors. After grouping together similar compilation errors, we found that they fell into 4 major error categories: *Undefined Identifier*, *Incorrect Argument passing*, *Incorrect Using Statements* and *Incorrect Return Type*. Upon looking at some instances of each category, we identified patterns of mistakes in the model's suggestions that cause the build to fail with these errors.

We noticed that the *Undefined Identifier* errors tend to happen when the model tries to use methods or classes outside of the provided context. As the model can only guess what other classes exist in the project and the methods contained within, it sometimes makes calls to methods that do not exist. We believe this could be improved by incorporating additional information regarding other classes within the project to the input, such as the ones that may be

used in the input code or from imported namespaces, along with the input context.

The *Incorrect Argument* errors also tend to occur when the model calls a method outside of provided the context. This results in the model passing in the wrong arguments types or number of arguments by making calls to method overloads that don't exist. We often saw this occur when the model tried to call member methods within some project-specific classes that were instantiated somewhere in the input code.

Cases for the *Incorrect Using Statements* follow a similar pattern as well. Here the model tries to import namespaces within the repo that don't exist or from packages that aren't in the build files. Since it doesn't know what other files exist in the project or the packages included in build, it often adds incorrect import statements.

The fourth category, *Type Mismatch*, occurs when the model suggests modifications that change the types of one or more class attributes, which get used elsewhere in the class. Since it can only modify the methods that are included in the input context (due to limited window), it is unable to modify these other methods. Other reasons for these errors include mismatch caused by changing the return type of a method, when the input class implements an interface, since changing the type would cause the method in the parent to not be overridden, leading to a compiler error.

Based on these observations, we believe a significant portion of above errors could be resolved by including a larger context containing more methods in the input class or even other classes/files in the project, through extended context [6]. We leave this exploration to future work.

**Table 6: Main reasons for compilation errors.**

Error Cause	Error Codes	Occurrences	% of Errors
Undefined Identifier	CS1061, CS0117, CS0246, CS0103, CS1579, etc.	3672	46.7
Incorrect Arguments	CS1503, CS1501, CS1729, CS7036, CS0305, CS0029, CS0019, etc.	2758	35.1
Incorrect Using Statements	CS0234	610	7.8
Type Mismatch	CS0266, CS0738, CS0508, etc.	246	3.1
Other Mistakes	CS0021, CS0122, (~120 misc. codes)	574	7.3
Total		7860	100.0

## 5.4 Running Benchmark Tests

The next step is to run benchmark tests for each of the changes that pass unit testing stage. However, before we run the benchmark tests we had to make some changes to the provided benchmark test suite to ensure the tests track the right metrics and that results are comparable among separate runs. By default, BenchmarkDotNet tests do not track allocations. For 22 out of the test repos, we found that memory tracking wasn't enabled and we had to enable it ourselves by adding a `[MemoryDiagnoser]` attribute to the class containing the benchmarks. Changing this does not affect the results for other metrics tracked by the benchmarks like test durations. Another change we had to make to make the numbers



comparable between separate runs was to add seeds to instances of random number generators instantiated in the benchmarking code. This is to ensure that the tests are deterministic so that the results can be compared between separate runs of the tests.

Additionally, to ensure no interference from background processes, we run the benchmark test in a sterile work environment with minimal workload other than the test itself. We first run the benchmark tests without any changes to measure the baseline performance of the application and then once after applying each of the changes that passed unit testing.

## 5.5 Analyzing Benchmark Results (RQ4)

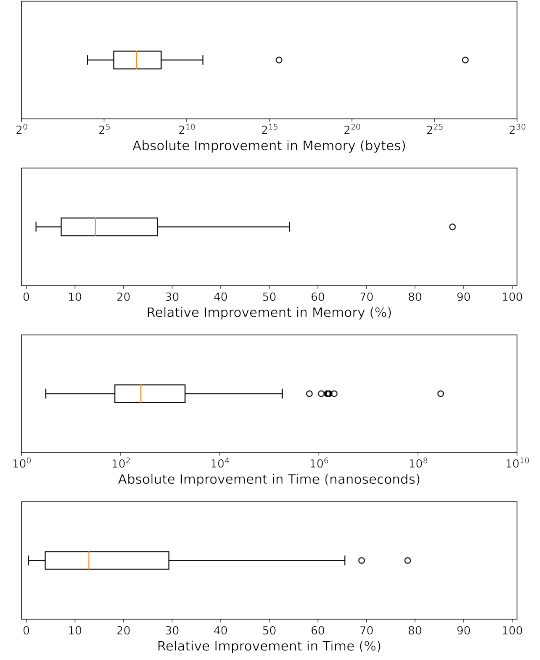
**5.5.1 Comparing Against Baseline.** Allocations are expected to stay consistent for C# applications as long as the benchmark tests are deterministic, so it is easy to tell if the change has improved memory usage by comparing the "Allocated" column (as shown in Figure 6). We consider a change to be a performance improvement in terms of Memory if it reduces allocations compared to the baseline.

For test duration, we use the "Mean", "StdDev" and "Iterations" columns, representing the sample mean, standard deviation and size, respectively. We make the assumption that the test duration readings are normally distributed. For each benchmark in the suggestion sample, we conduct a one-tailed Welch's  $t$ -test at 5% significance level to determine if the population mean of the suggestion code's sample is less than the population mean of the baseline (unmodified code) sample for the corresponding benchmark. In other words, our null hypothesis is that the population means of two samples are equal and the alternative hypothesis is that the population mean for suggestion is lower than baseline. We discard the suggestions that fail to reject the null-hypothesis. Following this, we conduct some additional checks on the remaining suggestions to reduce false positives and to ensure the change provides a significant enough improvement to be reported to the user. For this check, we use the "Q1" and "Q3" columns, which represent the first and third quartiles of the sample, respectively. We consider a suggestion to be a significant performance improvement over the baseline in terms of test durations, if the suggestion's upper Tukey fence is found to be lower than the the baseline's lower Tukey fence i.e. if  $Q_{3_{suggestion}} + 1.5(IQR_{suggestion}) < (Q_{1_{baseline}} - 1.5(IQR_{baseline}))$ , where IQR is the interquartile range,  $Q_3 - Q_1$ . Since there may be noise from background processes, this criteria also allows us to be robust to outliers and have fewer false positives. Finally, we also ensure that an improvement in allocation or test duration does not cause the other to deteriorate.

**5.5.2 Submitting a Perf Improvement PR.** Upon comparing the results against the baseline, we found that 543 suggestions improve performance metrics. These changes were saturated within 41 of the 201 methods. For each method, we verify up to 10 suggestions that pass unit tests and improve memory/test duration with a performance expert and submit a PR containing the first change that is a valid improvement. In case a project has correct suggestions for multiple methods, we squash all changes into a single PR.

For the cases where the model had generated correct suggestions, it was usually able to suggest the correct patch within the first or the second suggestion that passed unit tests and improved performance. Often times it suggested multiple distinct correct patches

that seemed to improve performance. Figure 1 shows two examples of valid performance improvement patches suggested during this evaluation that have been approved by the project owners.



**Figure 7: Above boxplots show improvement in benchmark test durations and allocations over baseline due to DeepPERF suggestions. Top two boxplots show the absolute and relative improvement over baseline allocations, respectively. Similarly, the bottom two show the absolute and relative improvement over baseline test durations.**

Figure 7 shows the improvement in benchmark durations and allocations due to DeepPERF's suggestions compared to baseline numbers. Looking at the relative improvement, we see that DeepPERF's suggestions typically provide ~15% improvement in terms of both allocations and test duration. Interestingly, for allocations, a few of our suggestions provide improvement on the order of KBs or even MBs. While others on the lower end do seem to veer in the territory of micro-optimizations, it should be noted that this is from benchmark tests and the developer may have simply written their benchmark test to be small (i.e. fewer iterations). We also don't know how often the tested code is run when the application sees use by a real customer. Depending on how often the code being tested is exercised during the application's runtime, especially if it appears on the application's hot-path, even these smaller improvements could improve performance significantly.

**RQ4:** In summary, we found that for 28 out of the 41 methods, DeepPERF had at least one correct performance improvement suggestion. We've submitted a total of 19 PRs, 9 of which have

already been approved by the project owners demonstrating the usefulness of our suggestions.

## 5.6 False Positives (RQ5)

One of our PRs was closed because the repo was not open to external contributions. However, the developer did not comment as to whether they considered the changes to be incorrect. Our remaining 8 PRs are still "Open" waiting for a response from the project owner. 13 out of the 41 methods DeepPERF found an optimization for turned out to be false positives i.e. they only had incorrect suggestions that seemed to improve benchmark results and somehow managed passed unit tests. This is a known issue in such models as they often generate suggestions that are test suite adequate, but otherwise turn out to be incorrect. While we make sure the methods we test have a high code coverage, that doesn't guarantee that the unit test will detect all mistakes as it may not be written to specifically test the particular method being modified. Another reason could be that the test suite itself is lacking. One way to combat these cases would be to generate additional unit cases and use them as further validation in addition to user-provided unit tests. One could also train an additional classifier to determine whether a change is correct and use it for filtration. We leave these explorations for future work.

**RQ5:** For a majority of methods, ~68%, that DeepPERF found improvements for, the changes were found to be valid and correctly passed the unit tests. However, in 13 out of 41 methods we only found invalid suggestions some of which were able to pass unit tests. While this is not insignificant, we plan to address this in future work.

## 6 THREATS TO VALIDITY

DeepPERF focuses on single file performance improvements, but often performance changes require modifications to multiple classes or even files. To generalize our approach to multiple file performance improvements, one could use techniques like extended context [6] to be able to pass in a larger context potentially from multiple classes or files. Another challenge when constructing input/output pairs from a given commit, would be determining which changes within the commit are related. A possible way to address this could be to by establishing caller-callee information between methods across files or use import statements to see which files are likely connected to the change. At inference time, one could generate the input based on caller-callee relationship among files and apply suggested changes to all the files involved.

Our "in-the-wild" evaluation used benchmark tests to validate the performance gains from our suggestions. It is difficult to know how much, if any, performance improvement this would provide to the end-user of the application. But, the fact that the developer wrote benchmark tests for these methods is a strong indication that the code must be frequently exercised and expected to be on the application's hot-path. Future work could replace benchmarking and combine our model with profiling or load testing instead to assess the performance gains in a more realistic usage scenario.

## 7 RELATED WORK

We describe how our work complements prior work in performance bug detection as well as automated bug detection.

### 7.1 Performance Bug Detection

There is a rich history of building tools for detecting performance bugs and improving performance. The majority of these tools identify code locations that take a long time to execute. Several tools generate or select tests for performance testing [33, 11, 3]. Other performance detection tools focus on detecting a specific type of performance bug. For instance, a set of tools have been developed for detecting runtime bloat [29, 31, 10], low-utility data structures [30], database related performance anti-patterns [4], false sharing problem in multi-threaded software [15], and detecting inefficient loops [17, 28]. Our tool extends the prior work on performance bug detection by developing a system that focuses on alleviating general performance problems and considers both source code features as well as performance symptoms through benchmarking.

### 7.2 Automatic Bug Detection

Prior work has investigated the use of static analyzers for detecting software bugs [32, 27, 22, 7]. More recently, researchers have started to explore the usage of machine learning for both software bug detection and bug fix. For instance, in C/C++, VulDeePecker uses deep learning to detect two types of vulnerabilities. Similarly, Russell et al. [21] propose a machine learning based method vulnerability detection in C/C++ code bases. In Java, Pang et al. [18] trained a machine learning model to predict static analyzer labels for Java source code. DeepFix leverages deep learning to generate fixes for simple syntax errors [12]. We are uniquely contributing to this area of research by leveraging neural networks for detecting optimization opportunities and suggesting performance improvements.

## 8 CONCLUSIONS AND FUTURE WORK

Detecting and fixing performance bugs remains an important yet challenging problem in the software development process. Our work makes three contributions to address this problem. First, we present a novel transformer based model to automatically generate patches providing performance improvement. Second, we conduct an empirical evaluation of our model to show that it outperforms the baselines over a dataset of performance optimizations collected from performance commits made by C# developers to repos on GitHub. Through this evaluation, we showed that our model is able to provide a wide-range of performance optimizations. We plan to release this dataset upon acceptance of this work as we believe that it will inform future studies in this space. Finally, we demonstrate a highly practical, end-to-end pipeline presenting our vision of automatically generating performance improvements for real world projects. This pipeline consists of our model alongside unit-testing and benchmarking, used to validate the generated patches. We also submit pull-requests containing the optimizations generated by this pipeline. Several of these PRs have already been merged, showing that our changes are considered valuable by the project owners.

## REFERENCES

- [1] Miltiadis Allamanis. “The adverse effects of code duplication in machine learning models of code”. In: Oct. 2019, pp. 143–153. ISBN: 978-1-4503-6995-4. doi: 10.1145/3359591.3359735.
- [2] Pavol Bielik, Veselin Raychev, and Martin Vechev. “Learning a static analyzer from data”. In: *International Conference on Computer Aided Verification*. Springer. 2017, pp. 233–253.
- [3] Jacob Burnim, Sudeep Juvekar, and Koushik Sen. “WISE: Automated Test Generation for Worst-Case Complexity”. In: *Proceedings of the 31st International Conference on Software Engineering*. ICSE ’09. USA: IEEE Computer Society, 2009, pp. 463–473. ISBN: 9781424434534. doi: 10.1109/ICSE.2009.5070545. URL: <https://doi.org/10.1109/ICSE.2009.5070545>.
- [4] Tse-Hsun Chen, Weiyei Shang, Zhen Ming Jiang, Ahmed E. Hassan, Mohamed Nasser, and Parminder Flora. “Detecting Performance Anti-Patterns for Applications Developed Using Object-Relational Mapping”. In: *Proceedings of the 36th International Conference on Software Engineering*. ICSE 2014. New York, NY, USA: Association for Computing Machinery, 2014, pp. 1001–1012. ISBN: 9781450327565.
- [5] Colin Clement, Dawn Drain, Jonathan Timcheck, Alexey Svyatkovskiy, and Neel Sundaresan. “PyMT5: Multi-mode Translation of Natural Language and Python Code with Transformers”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 9052–9065.
- [6] Colin Clement, Shuai Lu, Xiaoyu Liu, Michele Tufano, Dawn Drain, Nan Duan, Neel Sundaresan, and Alexey Svyatkovskiy. “Long-Range Modeling of Source Code Files with eWASH: Extended Window Access by Syntax Hierarchy”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4713–4722. doi: 10.18653/v1/2021.emnlp-main.387. URL: <https://aclanthology.org/2021.emnlp-main.387>.
- [7] Coverity. URL: <https://scan.coverity.com/>.
- [8] Daniel J. Dean, Hiep Nguyen, Xiaohui Gu, Hui Zhang, Junghwan Rhee, Nipun Arora, and Geoff Jiang. “PerfScope: Practical Online Server Performance Bug Inference in Production Cloud Computing Infrastructures”. In: *Proceedings of the ACM Symposium on Cloud Computing*. SOCC ’14. Seattle, WA, USA: Association for Computing Machinery, 2014, pp. 1–13. ISBN: 9781450332521. doi: 10.1145/2670979.2670987. URL: <https://doi.org/10.1145/2670979.2670987>.
- [9] Dawn Drain, Colin B. Clement, Guillermo Serrato, and Neel Sundaresan. “Deep-Debug: Fixing Python Bugs Using Stack Traces, Backtranslation, and Code Skeletons”. In: *ArXiv abs/2105.09352* (2021).
- [10] Bruno Dufour, Barbara G. Ryder, and Gary Sevitky. “A Scalable Technique for Characterizing the Usage of Temporaries in Framework-Intensive Java Applications”. In: *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. SIGSOFT ’08/FSE-16. Atlanta, Georgia: Association for Computing Machinery, 2008, pp. 59–70. ISBN: 9781595939951. doi: 10.1145/1453101.1453111. URL: <https://doi.org/10.1145/1453101.1453111>.
- [11] Mark Grechanik, Chen Fu, and Qing Xie. “Automatically Finding Performance Problems with Feedback-Directed Learning Software Testing”. In: *Proceedings of the 34th International Conference on Software Engineering*. ICSE ’12. Zurich, Switzerland: IEEE Press, 2012, pp. 156–166. ISBN: 9781467310673.
- [12] Rahul Gupta, Soham Pal, Aditya Kanade, and Shirish Shevade. “Deepfix: Fixing common c language errors by deep learning”. In: *Proceedings of the aaai conference on artificial intelligence*. Vol. 31. 1. 2017.
- [13] Milan Jovic, Andrea Adamoli, and Matthias Hauswirth. “Catch Me If You Can: Performance Bug Detection in the Wild”. In: *SIGPLAN Not.* 46.10 (Oct. 2011), pp. 155–170. ISSN: 0362-1340. doi: 10.1145/2076021.2048081. URL: <https://doi.org/10.1145/2076021.2048081>.
- [14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 7871–7880.
- [15] Tongping Liu and Emery D. Berger. “SHERIFF: Precise Detection and Automatic Mitigation of False Sharing”. In: *Proceedings of the 2011 ACM International Conference on Object Oriented Programming Systems Languages and Applications*. OOPSLA ’11. Portland, Oregon, USA: Association for Computing Machinery, 2011, pp. 3–18. ISBN: 9781450309400. doi: 10.1145/2048066.2048070. URL: <https://doi.org/10.1145/2048066.2048070>.
- [16] Sifei Luan, Di Yang, Celeste Barnaby, Koushik Sen, and Satish Chandra. “Aroma: Code recommendation via structural code search”. In: *Proceedings of the ACM on Programming Languages* 3.OOPSLA (2019), pp. 1–28.
- [17] Adrian Nistor, Tian Jiang, and Lin Tan. “Discovering, Reporting, and Fixing Performance Bugs”. In: *Proceedings of the 10th Working Conference on Mining Software Repositories*. MSR ’13. San Francisco, CA, USA: IEEE Press, 2013, pp. 237–246. ISBN: 9781467329361.
- [18] Yulei Pang, Xiaozhen Xue, and Akbar Siyami Namin. “Predicting vulnerable software components through n-gram analysis and statistical feature selection”. In: *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2015, pp. 543–548.
- [19] Paolo Pialorsi and Marco Russo. *Introducing microsoft® linq*. Microsoft Press, 2007.
- [20] Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, Duyu Tang, M. Zhou, Ambrosio Blanco, and Shuai Ma. “CodeBLEU: a Method for Automatic Evaluation of Code Synthesis”. In: *ArXiv abs/2009.10297* (2020).
- [21] Rebecca Russell, Louis Kim, Lei Hamilton, Tomo Lazovich, Jacob Harer, Onur Ozdemir, Paul Ellingwood, and Marc McConley. “Automated vulnerability detection in source code using deep representation learning”. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2018, pp. 757–762.
- [22] SonarQube. URL: <https://www.sonarqube.org/>.
- [23] Linhai Song and Shan Lu. “Performance Diagnosis for Inefficient Loops”. In: May 2017, pp. 370–380. doi: 10.1109/ICSE.2017.41.
- [24] Linhai Song and Shan Lu. “Statistical debugging for real-world performance problems”. In: *Proceedings of the 2014 ACM International Conference on Object Oriented Programming Systems Languages & Applications*. 2014, pp. 561–578.
- [25] Alexey Svyatkovskiy, Shao Kun Deng, Shengyu Fu, and Neel Sundaresan. “IntelliCode Compose: Code Generation Using Transformer”. In: *28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE ’20)*. May 2020. URL: <https://www.microsoft.com/en-us/research/publication/intellicode-compose-code-generation-using-transformer/>.
- [26] Michele Tufano, Dawn Drain, Alexey Svyatkovskiy, Shao Kun Deng, and Neel Sundaresan. “Unit Test Case Generation with Transformers”. In: *ArXiv abs/2009.05617* (2020).
- [27] John Viega, Jon-Thomas Bloch, Yoshi Kohno, and Gary McGraw. “ITS4: A static vulnerability scanner for C and C++ code”. In: *Proceedings 16th Annual Computer Security Applications Conference (ACSAC’00)*. IEEE. 2000, pp. 257–267.
- [28] Xusheng Xiao, Shi Han, Dongmei Zhang, and Tao Xie. “Context-Sensitive Delta Inference for Identifying Workload-Dependent Performance Bottlenecks”. In: *Proceedings of the 2013 International Symposium on Software Testing and Analysis*. ISSTA 2013. Lugano, Switzerland: Association for Computing Machinery, 2013, pp. 90–100. ISBN: 9781450321594. doi: 10.1145/2483760.2483784. URL: <https://doi.org/10.1145/2483760.2483784>.
- [29] Guoqing Xu, Matthew Arnold, Nick Mitchell, Atanas Rountev, and Gary Sevitky. “Go with the Flow: Profiling Copies to Find Runtime Bloat”. In: *Proceedings of the 30th ACM SIGPLAN Conference on Programming Language Design and Implementation*. PLDI ’09. Dublin, Ireland: Association for Computing Machinery, 2009, pp. 419–430. ISBN: 9781605583921. doi: 10.1145/1542476.1542523. URL: <https://doi.org/10.1145/1542476.1542523>.
- [30] Guoqing Xu, Nick Mitchell, Matthew Arnold, Atanas Rountev, Edith Schonberg, and Gary Sevitky. “Finding Low-Utility Data Structures”. In: *Proceedings of the 31st ACM SIGPLAN Conference on Programming Language Design and Implementation*. PLDI ’10. Toronto, Ontario, Canada: Association for Computing Machinery, 2010, pp. 174–186. ISBN: 9781450300193. doi: 10.1145/1806596.1806617. URL: <https://doi.org/10.1145/1806596.1806617>.
- [31] Guoqing Xu and Atanas Rountev. “Detecting Inefficiently-Used Containers to Avoid Bloat”. In: *Proceedings of the 31st ACM SIGPLAN Conference on Programming Language Design and Implementation*. PLDI ’10. Toronto, Ontario, Canada: Association for Computing Machinery, 2010, pp. 160–173. ISBN: 9781450300193. doi: 10.1145/1806596.1806616. URL: <https://doi.org/10.1145/1806596.1806616>.
- [32] Zhongxing Xu, Ted Kremenek, and Jian Zhang. “A memory model for static analysis of C programs”. In: *International Symposium On Leveraging Applications of Formal Methods, Verification and Validation*. Springer. 2010, pp. 535–548.
- [33] Pingyu Zhang, Sebastian Elbaum, and Matthew B. Dwyer. “Automatic Generation of Load Tests”. In: *Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering*. ASE ’11. USA: IEEE Computer Society, 2011, pp. 43–52. ISBN: 9781457716386. doi: 10.1109/ASE.2011.6100093. URL: <https://doi.org/10.1109/ASE.2011.6100093>.