Personalized Code Readability Assessment: Are We There Yet?

Antonio Vitale, Emanuela Guglielmi, Rocco Oliveto, Simone Scalabrino

ICPC 2025

The 33rd IEEE/ACM International Conference on Program Comprehension

Ottawa, ON, Canada – April 27 – 28, 2025

CODE READABILITY

```
for (int i=1;i<=100;i++) {
   int a=((528>>i%15-1)&1)*4;
   int b=((-2128340926>>(i%15)*2)&3)*4;
   System.out.println("FizzBuzz".substring(a,b)+(a==b?i:""));
}
```

```
for (int i=1;i<=100;i++) {
   String fizzBuzz = "";
   if (i % 3 == 0)
      fizzBuzz += "Fizz";

   if (i % 5 == 0)
      fizzBuzz += "Buzz";

   if (fizzBuzz.isEmpty())
      fizzBuzz += i;

   System.out.println(fizzBuzz);
}
```

CODE READABILITY


TSE SPECIAL ISSUE ON THE ISSTA 2008 BEST PAPERS

Learning a Metric for Code Readability

Raymond P.L. Buse, Westley R. Weimer

(Invited Paper)

2009


A Simpler Model of Software Readability

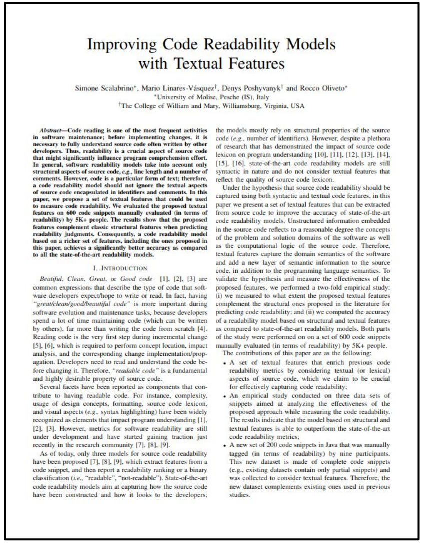
Daryl Posnett, University of California, Davis, Davis, CA, dpposnett@ucdavis.edu
Abram Hindle, University of California, Davis, Davis, CA, ah@softwareprocess.es
Prem Devanbu, University of California, Davis, Davis, CA, devanbu@ucdavis.edu

2011


A General Software Readability Model

Jonathan Dorn
Department of Computer Science
University of Virginia
Charlottesville, Virginia
jad5ju@virginia.edu

2012


Improving Code Readability Models with Textual Features

Simone Scalabrino*, Mario Linares-Vásquez†, Denys Poshyvanyk† and Rocco Oliveto*
*University of Molise, Pesche (IS), Italy
†The College of William and Mary, Williamsburg, Virginia, USA

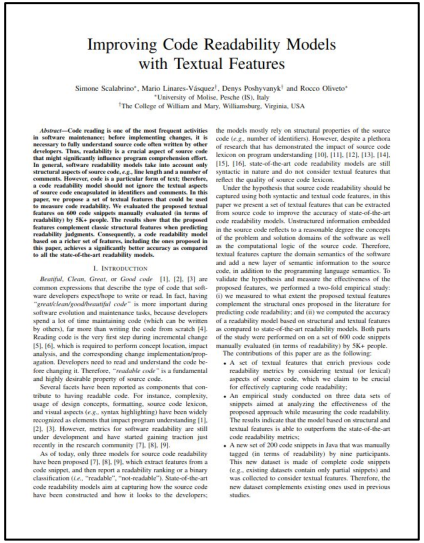
2016

CODE READABILITY

2009

2011

2012



2016

CODE READABILITY


TSE SPECIAL ISSUE ON THE ISSTA 2008 BEST PAPERS

Learning a Metric for Code Readability

Raymond P.L. Buse, Westley R. Weimer

(Invited Paper)

2009


A Simpler Model of Software Readability

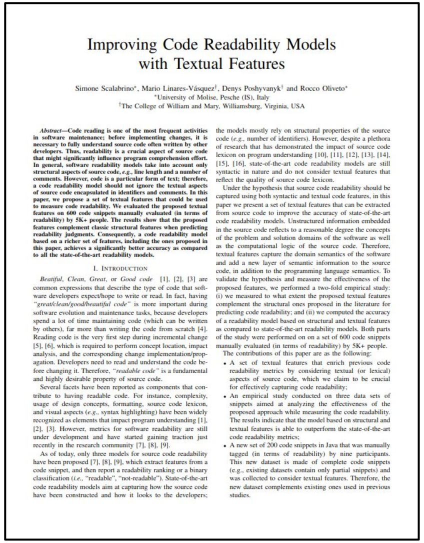
Daryl Posnett, University of California, Davis, Davis, CA, dpposnett@ucdavis.edu
Abram Hindle, University of California, Davis, Davis, CA, ah@softwareprocess.es
Prem Devanbu, University of California, Davis, Davis, CA, devanbu@ucdavis.edu

2011


A General Software Readability Model

Jonathan Dorn
Department of Computer Science
University of Virginia
Charlottesville, Virginia
jad5ju@virginia.edu

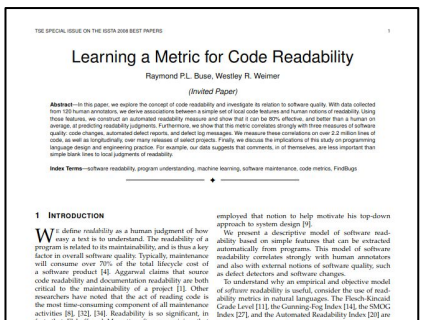
2012


Improving Code Readability Models with Textual Features

Simone Scalabrino*, Mario Linares-Vásquez†, Denys Poshyvanyk† and Rocco Oliveto*

2016

CODE READABILITY



TSE SPECIAL ISSUE ON THE ISSTA 2008 BEST PAPERS 1

Learning a Metric for Code Readability

Raymond P.L. Buse, Westley R. Weimer

(Invited Paper)

Abstract—In this paper, we explore the concept of code readability and investigate its relation to software quality. With data collected from 120 human annotators, we derive associations between a simple set of local code features and human notions of readability. Using those features, we construct an automated readability measure and show that it can be 80% effective, and better than a human on average, at predicting readability judgments. Furthermore, we show that this metric correlates strongly with three measures of software quality: code changes, automated defect reports, and defect log messages. We measure these correlations on over 2.2 million lines of code, as well as longitudinally, over many releases of select projects. Finally, we discuss the implications of this study on programming language design and engineering practice. For example, our data suggests that comments, in of themselves, are less important than simple blank lines to local judgments of readability.

Index Terms—software readability, program understanding, machine learning, software maintenance, code metrics, FindBugs

1 Introduction

"We define *readability* as a human judgment of how easy a text is to understand. The readability of a program is related to its maintainability, and is thus a key factor in overall software quality. Typically, maintenance will consume over 70% of the total lifecycle cost of a software product [4]. Aggarwal claims that source code readability and documentation readability are both critical to the maintainability of a project [1]. Other researchers have noted that the act of reading code is the most time-consuming component of all maintenance activities [8], [32], [34]. Readability is so significant, in fact, that Elshoff and Marcotty, after recognizing that many commercial programs were much more difficult to read than necessary, proposed adding a development phase in which the program is made more readable [10]. Knight and Myers suggested that one phase of software inspection should be a check of the source code for readability [21] to ensure maintainability, portability, and reusability of the code. Haneef proposed adding a dedicated readability and documentation group to the development team, observing that, "without established and consistent guidelines for readability, individual reviewers may not be able to help much" [15].

We hypothesize that programmers have some intuitive notion of this concept, and that program features such as indentation (e.g., as in Python [39]), choice of identifier names [33], and comments are likely to play a part. Dijkstra, for example, claimed that the readability of a program depends largely upon the simplicity of its sequencing control (e.g., he conjectured that `goto` unnecessarily complicates program understanding), and employed that notion to help motivate his top-down approach to system design [9].

We present a descriptive model of software readability based on simple features that can be extracted automatically from programs. This model of software readability correlates strongly with human annotators and also with external notions of software quality, such as defect detectors and software changes.

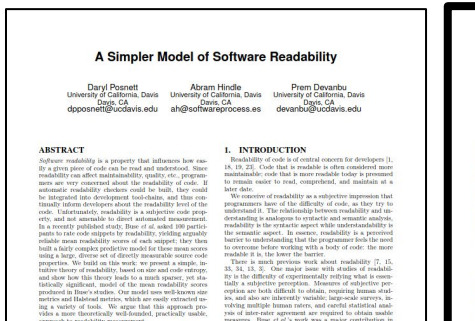
To understand why an empirical and objective model of *software* readability is useful, consider the use of readability metrics in natural languages. The Flesch-Kincaid Grade Level [11], the Gunning-Fog Index [14], the SMOG Index [27], and the Automated Readability Index [20] are just a few examples of readability metrics for ordinary text. These metrics are all based on simple factors such as average syllables per word and average sentence length. Despite this simplicity, they have each been shown to be quite useful in practice. Flesch-Kincaid, which has been in use for over 50 years, has not only been integrated into popular text editors including Microsoft Word, but has also become a United States governmental standard. Agencies, including the Department of Defense, require many documents and forms, internal and external, to meet have a Flesch readability grade of 10 or below (DOD MIL-M-38784B). Defense contractors also are often required to use it when they write technical manuals.

These metrics can help organizations gain some confidence that their documents meet goals for readability very cheaply, and have become ubiquitous for that reason. We believe that similar metrics, targeted specifically at source code and backed with empirical evidence for effectiveness, can serve an analogous purpose in the software domain. Readability metrics for the niche areas such as computer generated math [25], troff-input layout [3], and hypertext [16] have been found useful. We describe the first general readability metric for source code.

It is important to note that readability is not the same

• *Buse and Weimer are with the Department of Computer Science at The University of Virginia, Charlottesville, VA 22904. E-mail: {buse, weimer}@cs.virginia.edu*

2009



A Simpler Model of Software Readability

Daryl Posnett
University of California, Davis
Davis, CA
dpposnett@ucdavis.edu

Abram Hindle
University of California, Davis
Davis, CA
ah@softwareprocess.es

Prem Devanbu
University of California, Davis
Davis, CA
devanbu@ucdavis.edu

ABSTRACT

Software readability is a property that influences how easily a given piece of code can be read and understood. Since readability can affect maintainability, quality, etc., programmers are very concerned about the readability of code. If automatic readability checkers could be built, they could be integrated into development tool-chains, and thus continually inform developers about the readability level of the code. Unfortunately, readability is a subjective code property, and not amenable to direct automated measurement. In a recently published study, Buse *et al.* asked 100 participants to rate code snippets by readability, yielding arguably reliable mean readability scores of each snippet; they then built a fairly complex predictive model for these mean scores using a large, diverse set of directly measurable source code properties. We build on this work: we present a simple, intuitive theory of readability, based on size and code entropy, and show how this theory leads to a much sparser, yet statistically significant, model of the mean readability scores produced in Buse's studies. Our model uses well-known size metrics and Halstead metrics, which are easily extracted using a variety of tools. We argue that this approach provides a more theoretically well-founded, practically usable, approach to readability measurement.

Categories and Subject Descriptors

H.1 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Human Factors, Theory, Measurement

Keywords

Readability, Halstead, Entropy, Replication

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICSE '11, 21-MAY-2011, Waikiki, Honolulu , USA
Copyright 2011 ACM 978-1-4503-0575-7/11/05 ...$10.00.

1. INTRODUCTION

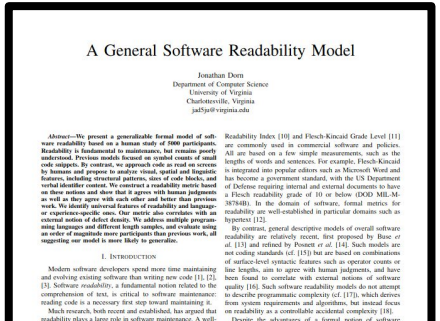
Readability of code is of central concern for developers [1, 18, 19, 23]. Code that is readable is often considered more maintainable; code that is more readable today is presumed to remain easier to read, comprehend, and maintain at a later date.

We conceive of readability as a subjective impression that programmers have of the difficulty of code, as they try to understand it. The relationship between readability and understanding is analogous to syntactic and semantic analysis; readability is the syntactic aspect while understandability is the semantic aspect. In essence, readability is a perceived barrier to understanding that the programmer feels the need to overcome before working with a body of code: the more readable it is, the lower the barrier.

There is much previous work about readability [7, 15, 35, 34, 13, 3]. One major issue with studies of readability is the difficulty of experimentally reifying what is essentially a subjective perception. Measures of subjective perception are both difficult to obtain, requiring human studies, and also are inherently variable; large-scale surveys, involving multiple human raters, and careful statistical analysis of inter-rater agreement are required to obtain usable measures. Buse *et al.*'s work was a major contribution in this area; they conducted a fairly large-scale study, asking human subjects to provide subjective rating scores of the readability of code snippets. Those scores were validated and aggregated to yield mean-opinion-scores [6]. The result of this extensive and time-consuming study was a set of code-snippets, $\mathcal{S}$, accompanied by mean subjective readability scores: $\mathcal{O}(s)$, $s \in \mathcal{S}$. Buse *et al.* then gathered direct, automatically-derived, token-level measures, $\mathcal{M}(s)$, of the code snippets. They then built a logistic regression model to predict the subjective, laboriously gathered scores, $\mathcal{O}(s)$, using the objective, automatically gathered direct measures, $\mathcal{M}(s)$. The $\mathcal{M}(s)$ were essentially token-level measures, but even so, were able to predict $\mathcal{O}(s)$ to some degree. This was an important contribution since it opens up the possibility of automatic tools that can provide readability scores as feedback to developers for every commit they make; this continuous feedback might potentially improve readability, and thus maintainability, of code over time.

In this paper we improve upon Buse *et al.*'s model, yielding a model that is *simpler, better performing* and *theoretically well-founded* in both classical software engineering and basic information theory. In particular we argue that fundamental aspects of readability have been measured since the 1970s by Halstead's software science metrics [21]. We show that the Halstead metrics can be used to improve upon the

2011



A General Software Readability Model

Jonathan Dorn
Department of Computer Science
University of Virginia
Charlottesville, Virginia
jad5ya@virginia.edu

***Abstract*—We present a generalizable formal model of software readability based on a human study of 5000 participants. Readability is fundamental to maintenance, but remains poorly understood. Previous models focused on symbol counts of small code snippets. By contrast, we approach code as read on screens by humans and propose to analyze visual, spatial and linguistic features, including structural patterns, sizes of code blocks, and verbal identifier content. We construct a readability metric based on these notions and show that it agrees with human judgments as well as they agree with each other and better than previous work. We identify universal features of readability and language- or experience-specific ones. Our metric also correlates with an external notion of defect density. We address multiple programming languages and different length samples, and evaluate using an order of magnitude more participants than previous work, all suggesting our model is more likely to generalize.**

I. Introduction

Modern software developers spend more time maintaining and evolving existing software than writing new code [1], [2], [3]. Software *readability*, a fundamental notion related to the comprehension of text, is critical to software maintenance: reading code is a necessary first step toward maintaining it.

Much research, both recent and established, has argued that readability plays a large role in software maintenance. A well-known example is Knuth, who viewed readability as essential to his notion of Literate Programming [4]. He argued that a program should be viewed as "a piece of literature, addressed to human beings" and that a readable program is "more robust, more portable, [and] more easily maintained". Haneef argued in favor of a development group dedicated to readability and documentation: "without established and consistent guidelines for readability, individual reviewers may not be able to help much" [5]. Knight and Myers argued that a source-level check for readability improves portability, maintainability and reusability and should thus be a first-class phase of software inspection [6]. Basili *et al.* showed that inspections guided by reading techniques are better at revealing defects [7]. An entire development phase aimed at improving readability was proposed by Elshoff and Marcotty, who observed that many commercial programs were unnecessarily difficult to read [8]. More recently, a 2012 survey of over 100 developers and managers at Microsoft by Buse and Zimmermann found that 90% of responders desire readability as a software analytic feature, placing it among the top three in their survey [9].

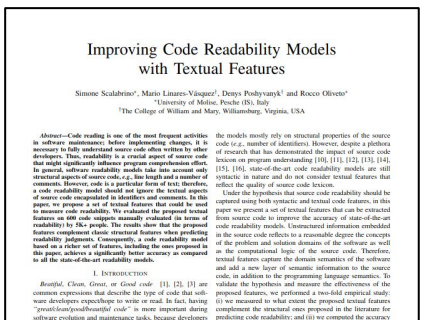
Readability metrics are well-established in the domain of non-software natural language. Metrics such as the Automated Readability Index [10] and Flesch-Kincaid Grade Level [11] are commonly used in commercial software and policies. All are based on a few simple measurements, such as the lengths of words and sentences. For example, Flesch-Kincaid is integrated into popular editors such as Microsoft Word and has become a government standard, with the US Department of Defense requiring internal and external documents to have a Flesch readability grade of 10 or below (DOD MIL-M-38784B). In the domain of software, formal metrics for readability are well-established in particular domains such as hypertext [12].

By contrast, general descriptive models of overall software readability are relatively recent, first proposed by Buse *et al.* [13] and refined by Posnett *et al.* [14]. Such models are not coding standards (cf. [15]) but are based on combinations of surface-level syntactic features such as operator counts or line lengths, aim to agree with human judgments, and have been found to correlate with external notions of software quality [16]. Such software readability models do not attempt to describe programmatic complexity (cf. [17]), which derives from system requirements and algorithms, but instead focus on readability as a controllable accidental complexity [18].

Despite the advantages of a formal notion of software readability, previous readability metrics do not adequately *generalize*. They are based on small (typically 7-line) snippets of code from a single programming language, are tied to shallow surface features that do not account for visual presentation or linguistic meaning, and derive from the judgments of a relatively small number of students [13], [14]. We propose a readability model that addresses all of these concerns while remaining lightweight and applicable.

Intuitively, the effectiveness of syntax highlighting suggests that visual or geometric formatting significantly impacts code readability. Similarly, the prevalence of variable naming standards (e.g., underscores, camel case, Hungarian notation) suggests that meaningful linguistic information is captured by identifiers. We thus propose the first incorporation of geometric, pattern-based and linguistic aspects and features into an automated readability metric. For example, code in which the "=" operators in a sequence of assignment statements "line up" vertically on the screen may be viewed as more readable, as may code in which identifiers contain English synonyms or code in which comments form a colored rectangular block. We propose to incorporate such features into our model of readability.

2012



Improving Code Readability Models with Textual Features

Simone Scalabrino*, Mario Linares-Vásquez†, Denys Poshyvanyk‡ and Rocco Oliveto*
*University of Molise, Pesche (IS), Italy
†The College of William and Mary, Williamsburg, Virginia, USA

***Abstract*—Code reading is one of the most frequent activities in software maintenance; therefore, improving changes, it is necessary to fully understand source code often written by other developers. Thus, readability is a crucial aspect of source code that might significantly influence program comprehension efforts. In general, software readability models take into account only structural aspects of source code, e.g., line length and a number of comments. However, code is a particular form of text; therefore, a code readability model should not ignore the textual aspects of source code encapsulated in identifiers and comments. In this paper, we propose a set of textual features that could be used to measure code readability. We evaluated the proposed textual features on 600 code snippets manually evaluated (in terms of readability) by 5K+ people. The results show that the proposed features complement classic structural features when predicting readability judgments. Consequently, a code readability model based on a richer set of features, including the ones proposed in this paper, achieves a significantly better accuracy as compared to all the state-of-the-art readability models.**

I. Introduction

Beautiful, Clean, Great, or *Good code* [1], [2], [3] are common expressions that describe the type of code that software developers expect/hope to write or read. In fact, having *"great/clean/good/beautiful code"* is more important during software evolution and maintenance tasks, because developers spend a lot of time maintaining code (which can be written by others), far more than writing the code from scratch [4]. Reading code is the very first step during incremental change [5], [6], which is required to perform concept location, impact analysis, and the corresponding change implementation/propagation. Developers need to read and understand the code before changing it. Therefore, *"readable code"* is a fundamental and highly desirable property of source code.

Several facets have been reported as components that contribute to readable source code. For instance, complexity, usage of design concepts, formatting, source code lexicon, and visual aspects (*e.g.*, syntax highlighting) have been widely recognized as elements that impact program understanding [1], [2], [3]. However, metrics for software readability are still under development and have started gaining traction just recently in the research community [7], [8], [9].

As of today, only three models for source code readability have been proposed [7], [8], [9], which extract features from a code snippet, and then report a readability ranking or a binary classification (*i.e.*, "readable", "not-readable"). State-of-the-art code readability models aim at capturing how the source code have been constructed and how it looks to the developers; the models mostly rely on structural properties of the source code (*e.g.*, number of identifiers). However, despite a plethora of research that has demonstrated the impact of source code lexicon on program understanding [10], [11], [12], [13], [14], [15], [16], state-of-the-art code readability models are still syntactic in nature and do not consider textual features that reflect the quality of source code lexicon.

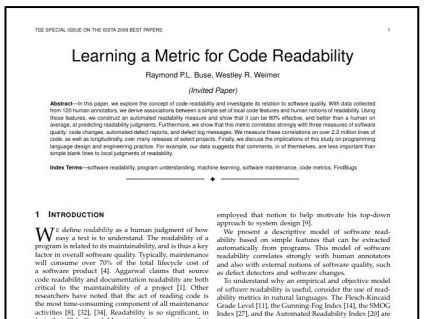
Under the hypothesis that source code readability should be captured using both syntactic and textual code features, in this paper we present a set of textual features that can be extracted from source code to improve the accuracy of state-of-the-art code readability models. Unstructured information embedded in the source code reflects to a reasonable degree the concepts of the problem and solution domains of the software as well as the computational logic of the source code. Therefore, textual features capture the domain semantics of the software and add a new layer of semantic information to the source code, in addition to the programming language semantics. To validate the hypothesis and measure the effectiveness of the proposed features, we performed a two-fold empirical study: (i) we measured to what extent the proposed textual features complement the structural ones proposed in the literature for predicting code readability; and (ii) we computed the accuracy of a readability model based on structural and textual features as compared to state-of-the-art readability models. Both parts of the study were performed on on a set of 600 code snippets manually evaluated (in terms of readability) by 5K+ people.

The contributions of this paper are as the following:

- A set of textual features that enrich previous code readability metrics by considering textual (or lexical) aspects of source code, which we claim to be crucial for effectively capturing code readability;
- An empirical study conducted on three data sets of snippets aimed at analyzing the effectiveness of the proposed approach while measuring the code readability. The results indicate that the model based on structural and textual features is able to outperform the state-of-the-art code readability metrics;
- A new set of 200 code snippets in Java that was manually tagged (in terms of readability) by nine participants. This new dataset is made of complete code snippets (e.g., existing datasets contain only partial snippets) and was collected to consider textual features. Therefore, the new dataset complements existing ones used in previous studies.

2016

CODE READABILITY



TSE SPECIAL ISSUE ON THE ISSTA 2008 BEST PAPERS 1

Learning a Metric for Code Readability

Raymond P.L. Buse, Westley R. Weimer

(Invited Paper)

Abstract—In this paper, we explore the concept of code readability and investigate its relation to software quality. With data collected from 120 human annotators, we derive associations between a simple set of local code features and human notions of readability. Using those features, we construct an automated readability measure and show that it can be 80% effective, and better than a human on average, at predicting readability judgments. Furthermore, we show that this metric correlates strongly with three measures of software quality: code changes, automated defect reports, and defect log messages. We measure these correlations on over 2.2 million lines of code, as well as longitudinally, over many releases of select projects. Finally, we discuss the implications of this study on programming language design and engineering practice. For example, our data suggests that comments, in of themselves, are less important than simple blank lines to local judgments of readability.

Index Terms—software readability, program understanding, machine learning, software maintenance, code metrics, FindBugs

1 Introduction

"We define *readability* as a human judgment of how easy a text is to understand. The readability of a program is related to its maintainability, and is thus a key factor in overall software quality. Typically, maintenance will consume over 70% of the total lifecycle cost of a software product [4]. Aggarwal claims that source code readability and documentation readability are both critical to the maintainability of a project [1]. Other researchers have noted that the act of reading code is the most time-consuming component of all maintenance activities [8], [32], [34]. Readability is so significant, in fact, that Elshoff and Marcotty, after recognizing that many commercial programs were much more difficult to read than necessary, proposed adding a development phase in which the program is made more readable [10]. Knight and Myers suggested that one phase of software inspection should be a check of the source code for readability [21] to ensure maintainability, portability, and reusability of the code. Haneef proposed adding a dedicated readability and documentation group to the development team, observing that, "without established and consistent guidelines for readability, individual reviewers may not be able to help much" [15].

We hypothesize that programmers have some intuitive notion of this concept, and that program features such as indentation (e.g., as in Python [39]), choice of identifier names [33], and comments are likely to play a part. Dijkstra, for example, claimed that the readability of a program depends largely upon the simplicity of its sequencing control (e.g., he conjectured that `goto` unnecessarily complicates program understanding), and employed that notion to help motivate his top-down approach to system design [9].

We present a descriptive model of software readability based on simple features that can be extracted automatically from programs. This model of software readability correlates strongly with human annotators and also with external notions of software quality, such as defect detectors and software changes.

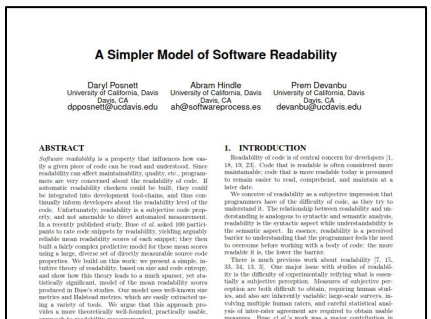
To understand why an empirical and objective model of *software* readability is useful, consider the use of readability metrics in natural languages. The Flesch-Kincaid Grade Level [11], the Gunning-Fog Index [14], the SMOG Index [27], and the Automated Readability Index [20] are just a few examples of readability metrics for ordinary text. These metrics are all based on simple factors such as average syllables per word and average sentence length. Despite this simplicity, they have each been shown to be quite useful in practice. Flesch-Kincaid, which has been in use for over 50 years, has not only been integrated into popular text editors including Microsoft Word, but has also become a United States governmental standard. Agencies, including the Department of Defense, require many documents and forms, internal and external, to meet have a Flesch readability grade of 10 or below (DOD MIL-M-38784B). Defense contractors also are often required to use it when they write technical manuals.

These metrics can help organizations gain some confidence that their documents meet goals for readability very cheaply, and have become ubiquitous for that reason. We believe that similar metrics, targeted specifically at source code and backed with empirical evidence for effectiveness, can serve an analogous purpose in the software domain. Readability metrics for the niche areas such as computer generated math [25], freemap layout [3], and hypertext [16] have been found useful. We describe the first general readability metric for source code.

It is important to note that readability is not the same

• *Buse and Weimer are with the Department of Computer Science at The University of Virginia, Charlottesville, VA 22904. E-mail: {buse, weimer}@cs.virginia.edu*

2009



A Simpler Model of Software Readability

Daryl Posnett
University of California, Davis
Davis, CA
dpposnett@ucdavis.edu

Abram Hindle
University of California, Davis
Davis, CA
ah@softwareprocess.es

Prem Devanbu
University of California, Davis
Davis, CA
devanbu@ucdavis.edu

ABSTRACT

Software readability is a property that influences how easily a given piece of code can be read and understood. Since readability can affect maintainability, quality, etc., programmers are very concerned about the readability of code. If automatic readability checkers could be built, they could be integrated into development tool-chains, and thus continually inform developers about the readability level of the code. Unfortunately, readability is a subjective code property, and not amenable to direct automated measurement. In a recently published study, Buse *et al.* asked 100 participants to rate code snippets by readability, yielding arguably reliable mean readability scores of each snippet; they then built a fairly complex predictive model for these mean scores using a large, diverse set of directly measurable source code properties. We build on this work: we present a simple, intuitive theory of readability, based on size and code entropy, and show how this theory leads to a much sparser, yet statistically significant, model of the mean readability scores produced in Buse's studies. Our model uses well-known size metrics and Halstead metrics, which are easily extracted using a variety of tools. We argue that this approach provides a more theoretically well-founded, practically usable, approach to readability measurement.

Categories and Subject Descriptors

H.1 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Human Factors, Theory, Measurement

Keywords

Readability, Halstead, Entropy, Replication

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ISSTA '11, 21-MAY-2011, Waikiki, Honolulu , USA
Copyright 2011 ACM 978-1-4503-0571-7/11/05 ...$10.00.

1. INTRODUCTION

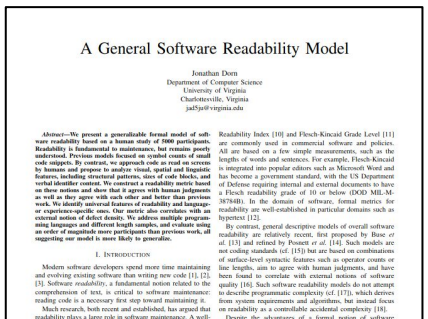
Readability of code is of central concern for developers [1, 18, 19, 23]. Code that is readable is often considered more maintainable; code that is more readable today is presumed to remain easier to read, comprehend, and maintain at a later date.

We conceive of readability as a subjective impression that programmers have of the difficulty of code, as they try to understand it. The relationship between readability and understanding is analogous to syntactic and semantic analysis; readability is the syntactic aspect while understandability is the semantic aspect. In essence, readability is a perceived barrier to understanding that the programmer feels the need to overcome before working with a body of code: the more readable it is, the lower the barrier.

There is much previous work about readability [7, 15, 35, 34, 13, 3]. One major issue with studies of readability is the difficulty of experimentally reifying what is essentially a subjective perception. Measures of subjective perception are both difficult to obtain, requiring human studies, and also are inherently variable; large-scale surveys, involving multiple human raters, and careful statistical analysis of inter-rater agreement are required to obtain usable measures. Buse *et al.*'s work was a major contribution in this area; they conducted a fairly large-scale study, asking human subjects to provide subjective rating scores of the readability of code snippets. Those scores were validated and aggregated to yield mean-opinion-scores [6]. The result of this extensive and time-consuming study was a set of code-snippets, $\mathcal{S}$, accompanied by mean subjective readability scores: $\mathcal{O}(s)$, $s \in \mathcal{S}$. Buse *et al.* then gathered direct, automatically-derived, token-level measures, $\mathcal{M}(s)$, of the code snippets. They then built a logistic regression model to predict the subjective, laboriously gathered scores, $\mathcal{O}(s)$, using the objective, automatically gathered direct measures, $\mathcal{M}(s)$. The $\mathcal{M}(s)$ were essentially token-level measures, but even so, were able to predict $\mathcal{O}(s)$ to some degree. This was an important contribution since it opens up the possibility of automatic tools that can provide readability scores as feedback to developers for every commit they make; this continuous feedback might potentially improve readability, and thus maintainability, of code over time.

In this paper we improve upon Buse *et al.*'s model, yielding a model that is *simpler, better performing* and *theoretically well-founded* in both classical software engineering and basic information theory. In particular we argue that fundamental aspects of readability have been measured since the 1970s by Halstead's software science metrics [21]. We show that the Halstead metrics can be used to improve upon the

2011



A General Software Readability Model

Jonathan Dorn
Department of Computer Science
University of Virginia
Charlottesville, Virginia
jad5ya@virginia.edu

***Abstract*—We present a generalizable formal model of software readability based on a human study of 5000 participants. Readability is fundamental to maintenance, but remains poorly understood. Previous models focused on symbol counts of small code snippets. By contrast, we approach code as read on screens by humans and propose to analyze visual, spatial and linguistic features, including structural patterns, sizes of code blocks, and verbal identifier content. We construct a readability metric based on these notions and show that it agrees with human judgments as well as they agree with each other and better than previous work. We identify universal features of readability and language- or experience-specific ones. Our metric also correlates with an external notion of defect density. We address multiple programming languages and different length samples, and evaluate using an order of magnitude more participants than previous work, all suggesting our model is more likely to generalize.**

I. Introduction

Modern software developers spend more time maintaining and evolving existing software than writing new code [1], [2], [3]. Software *readability*, a fundamental notion related to the comprehension of text, is critical to software maintenance: reading code is a necessary first step toward maintaining it.

Much research, both recent and established, has argued that readability plays a large role in software maintenance. A well-known example is Knuth, who viewed readability as essential to his notion of Literate Programming [4]. He argued that a program should be viewed as "a piece of literature, addressed to human beings" and that a readable program is "more robust, more portable, [and] more easily maintained". Haneef argued in favor of a development group dedicated to readability and documentation: "without established and consistent guidelines for readability, individual reviewers may not be able to help much" [5]. Knight and Myers argued that a source-level check for readability improves portability, reusability and reusability and should thus be a first-class phase of software inspection [6]. Basili *et al.* showed that inspections guided by reading techniques are better at revealing defects [7]. An entire development phase aimed at improving readability was proposed by Elshoff and Marcotty, who observed that many commercial programs were unnecessarily difficult to read [8]. More recently, a 2012 survey of over 100 developers and managers at Microsoft by Buse and Zimmermann found that 90% of responders desire readability as a software analytic feature, placing it among the top three in their survey [9].

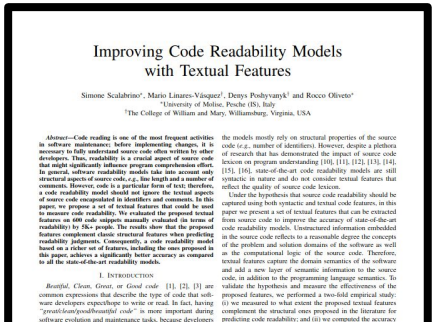
Readability metrics are well-established in the domain of non-software natural language. Metrics such as the Automated Readability Index [10] and Flesch-Kincaid Grade Level [11] are commonly used in commercial software and policies. All are based on a few simple measurements, such as the lengths of words and sentences. For example, Flesch-Kincaid is integrated into popular editors such as Microsoft Word and has become a government standard, with the US Department of Defense requiring internal and external documents to have a Flesch readability grade of 10 or below (DOD MIL-M-38784B). In the domain of software, formal metrics for readability are well-established in particular domains such as hypertext [12].

By contrast, general descriptive models of overall software readability are relatively recent, first proposed by Buse *et al.* [13] and refined by Posnett *et al.* [14]. Such models are not coding standards (cf. [15]) but are based on combinations of surface-level syntactic features such as operator counts or line lengths, aim to agree with human judgments, and have been found to correlate with external notions of software quality [16]. Such software readability models do not attempt to describe programmatic complexity (cf. [17]), which derives from system requirements and algorithms, but instead focus on readability as a controllable accidental complexity [18].

Despite the advantages of a formal notion of software readability, previous readability metrics do not adequately *generalize*. They are based on small (typically 7-line) snippets of code from a single programming language, are tied to shallow surface features that do not account for visual presentation or linguistic meaning, and derive from the judgments of a relatively small number of students [13], [14]. We propose a readability model that addresses all of these concerns while remaining lightweight and applicable.

Intuitively, the effectiveness of syntax highlighting suggests that visual or geometric formatting significantly impacts code readability. Similarly, the prevalence of variable naming standards (e.g., underscores, camel case, Hungarian notation) suggests that meaningful linguistic information is captured by identifiers. We thus propose the first incorporation of geometric, pattern-based and linguistic aspects and features into an automated readability metric. For example, code in which the "=" operators in a sequence of assignment statements "line up" vertically on the screen may be viewed as more readable, as may code in which identifiers contain English synonyms or code in which comments form a colored rectangular block. We propose to incorporate such features into our model of readability.

2012



Improving Code Readability Models with Textual Features

Simone Scalabrino*, Mario Linares-Vásquez†, Denys Poshyvanyk‡ and Rocco Oliveto*
*University of Molise, Pesche (IS), Italy
†The College of William and Mary, Williamsburg, Virginia, USA

***Abstract*—Code reading is one of the most frequent activities in software maintenance; before implementing changes, it is necessary to fully understand source code often written by other developers. Thus, readability is a crucial aspect of source code that might significantly influence program comprehension efforts. In general, software readability models take into account only structural aspects of source code, e.g., line length and a number of comments. However, code is a particular form of text; therefore, a code readability model should not ignore the textual aspects of source code encapsulated in identifiers and comments. In this paper, we propose a set of textual features that could be used to measure code readability. We evaluated the proposed textual features on 600 code snippets manually evaluated (in terms of readability) by 5K+ people. The results show that the proposed features complement classic structural features when predicting readability judgments. Consequently, a code readability model based on a richer set of features, including the ones proposed in this paper, achieves a significantly better accuracy as compared to all the state-of-the-art readability models.**

I. Introduction

Beautiful, Clean, Great, or *Good code* [1], [2], [3] are common expressions that describe the type of code that software developers expect/hope to write or read. In fact, having *"great/clean/good/beautiful code"* is more important during software evolution and maintenance tasks, because developers spend a lot of time maintaining code (which can be written by others), far more than writing the code from scratch [4]. Reading code is the very first step during incremental change [5], [6], which is required to perform concept location, impact analysis, and the corresponding change implementation/propagation. Developers need to read and understand the code before changing it. Therefore, *"readable code"* is a fundamental and highly desirable property of source code.

Several facets have been reported as components that contribute to readable source code. For instance, complexity, usage of design concepts, formatting, source code lexicon, and visual aspects (*e.g.*, syntax highlighting) have been widely recognized as elements that impact program understanding [1], [2], [3]. However, metrics for software readability are still under development and have started gaining traction just recently in the research community [7], [8], [9].

As of today, only three models for source code readability have been proposed [7], [8], [9], which extract features from a code snippet, and then report a readability ranking or a binary classification (*i.e.*, "readable", "not-readable"). State-of-the-art code readability models aim at capturing how the source code have been constructed and how it looks to the developers; the models mostly rely on structural properties of the source code (*e.g.*, number of identifiers). However, despite a plethora of research that has demonstrated the impact of source code lexicon on program understanding [10], [11], [12], [13], [14], [15], [16], state-of-the-art code readability models are still syntactic in nature and do not consider textual features that reflect the quality of source code lexicon.

Under the hypothesis that source code readability should be captured using both syntactic and textual code features, in this paper we present a set of textual features that can be extracted from source code to improve the accuracy of state-of-the-art code readability models. Unstructured information embedded in the source code reflects to a reasonable degree the concepts of the problem and solution domains of the software as well as the computational logic of the source code. Therefore, textual features capture the domain semantics of the software and add a new layer of semantic information to the source code, in addition to the programming language semantics. To validate the hypothesis and measure the effectiveness of the proposed features, we performed a two-fold empirical study: (i) we measured to what extent the proposed textual features complement the structural ones proposed in the literature for predicting code readability; and (ii) we computed the accuracy of a readability model based on structural and textual features as compared to state-of-the-art readability models. Both parts of the study were performed on on a set of 600 code snippets manually evaluated (in terms of readability) by 5K+ people.

The contributions of this paper are as the following:

- A set of textual features that enrich previous code readability metrics by considering textual (or lexical) aspects of source code, which we claim to be crucial for effectively capturing code readability.
- An empirical study conducted on three data sets of snippets aimed at analyzing the effectiveness of the proposed approach while measuring the code readability. The results indicate that the model based on structural and textual features is able to outperform the state-of-the-art code readability metrics;
- A new set of 200 code snippets in Java that was manually tagged (in terms of readability) by nine participants. This new dataset is made of complete code snippets (e.g., existing datasets contain only partial snippets) and was collected to consider textual features. Therefore, the new dataset complements existing ones used in previous studies.

2016

CODE READABILITY

TSE SPECIAL ISSUE ON THE ISSTA 2008 BEST PAPERS 1

Learning a Metric for Code Readability

Raymond P.L. Buse, Westley R. Weimer

(Invited Paper)

Abstract—In this paper, we explore the concept of code readability and investigate its relation to software quality. With data collected from 120 human annotators, we derive associations between a simple set of local code features and human notions of readability. Using those features, we construct an automated readability measure and show that it can be 80% effective, and better than a human on average, at predicting readability judgments. Furthermore, we show that this metric correlates strongly with three measures of software quality: code changes, automated defect reports, and defect log messages. We measure these correlations on over 2.2 million lines of code, as well as longitudinally, over many releases of select projects. Finally, we discuss the implications of this study on programming language design and engineering practice. For example, our data suggests that comments, in of themselves, are less important than simple blank lines to local judgments of readability.

Index Terms—software readability, program understanding, machine learning, software maintenance, code metrics, FindBugs

1 INTRODUCTION

We define *readability* as a human judgment of how easy a text is to understand. The readability of a program is related to its maintainability, and is thus a key factor in overall software quality. Typically, maintenance will consume over *70%* of the total lifecycle cost of a software product [4]. Aggarwal claims that source code readability and documentation readability are both critical to the maintainability of a project [1]. Other researchers have noted that the act of reading code is the most time-consuming component of all maintenance activities [8], [32], [34]. Readability is so significant, in fact, that Elshoff and Marcotty, after recognizing that many commercial programs were much more difficult to read than necessary, proposed adding a development phase in which the program is made more readable [10]. Knight and Myers suggested that one phase of software inspection should be a check of the source code for readability [21] to ensure maintainability, portability, and reusability of the code. Haneef proposed adding a dedicated readability and documentation group to the development team, observing that, "without established and consistent guidelines for readability, individual reviewers may not be able to help much" [15].

We hypothesize that programmers have some intuitive notion of this concept, and that program features such indentation (e.g., as in Python [39]), choice of identifier names [33], and comments are likely to play a part. Dijkstra, for example, claimed that the readability of a program depends largely upon the simplicity of its sequencing control (e.g., he conjectured that `goto` unnecessarily complicates program understanding), and employed that notion to help motivate his top-down approach to system design [9].

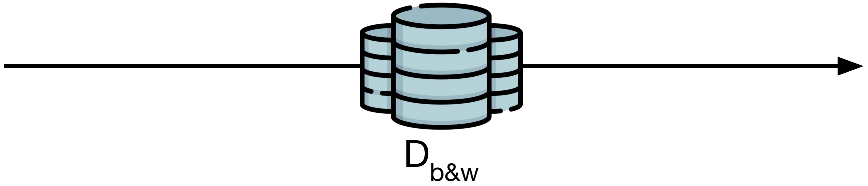
We present a descriptive model of software readability based on simple features that can be extracted automatically from programs. This model of software readability correlates strongly with human annotators and also with external notions of software quality, such as defect detectors and software changes.

To understand why an empirical and objective model of *software* readability is useful, consider the use of readability metrics in natural languages. The Flesch-Kincaid Grade Level [11], the Gunning-Fog Index [14], the SMOG Index [27], and the Automated Readability Index [20] are just a few examples of readability metrics for ordinary text. These metrics are all based on simple factors such as average syllables per word and average sentence length. Despite this simplicity, they have each been shown to be quite useful in practice. Flesch-Kincaid, which has been in use for over 50 years, has not only been integrated into popular text editors including Microsoft Word, but has also become a United States governmental standard. Agencies, including the Department of Defense, require many documents and forms, internal and external, to meet have a Flesch readability grade of 10 or below (DOD MIL-M-38784B). Defense contractors also are often required to use it when they write technical manuals.

These metrics can help organizations gain some confidence that their documents meet goals for readability very cheaply, and have become ubiquitous for that reason. We believe that similar metrics, targeted specifically at source code and backed with empirical evidence for effectiveness, can serve an analogous purpose in the software domain. Readability metrics for the niche areas such as computer generated math [25], treemap layout [3], and hypertext [16] have been found useful. We describe the first general readability metric for source code.

It is important to note that readability is not the same

• *Buse and Weimer are with the Department of Computer Science at The University of Virginia, Charlottesville, VA 22904. E-mail: {buse, weimer}@cs.virginia.edu*



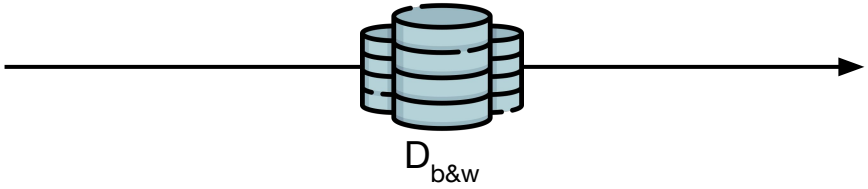
Snippets: 100
Developers: 120
Granularity: Snippet
Assessments: Complete

CODE READABILITY

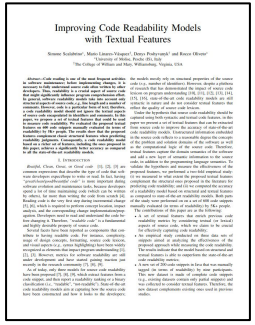
Learning a Metric for Code Readability

Raymond P.L. Buse, Westley R. Weimer

(Invited Paper)



Snippets: 100
Developers: 120
Granularity: Snippet
Assessments: Complete

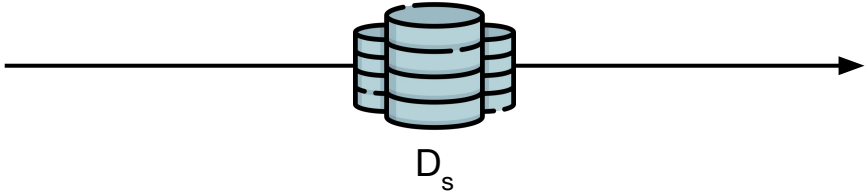


Improving Code Readability Models with Textual Features

Simone Scalabrino*, Mario Linares-Vásquez*, Denys Poshyvanyk†, and Rocco Oliveto*

*University of Molise, Pesche (IS), Italy

†The College of William and Mary, Williamsburg, Virginia, USA



Snippets: 200
Developers: 9
Granularity: Method
Assessments: Complete

CODE READABILITY

Developers Assessments

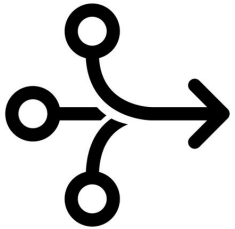
Snippets					
	3	4	2	4	5
	1	3	2	1	1
	4	1	5	4	5
	5	1	1	1	3
	4	3	2	5	5

CODE READABILITY

Developers Assessments

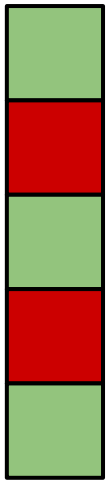
Snippets

3	4	2	4	5
1	3	2	1	1
4	1	5	4	5
5	1	1	1	3
4	3	2	5	5

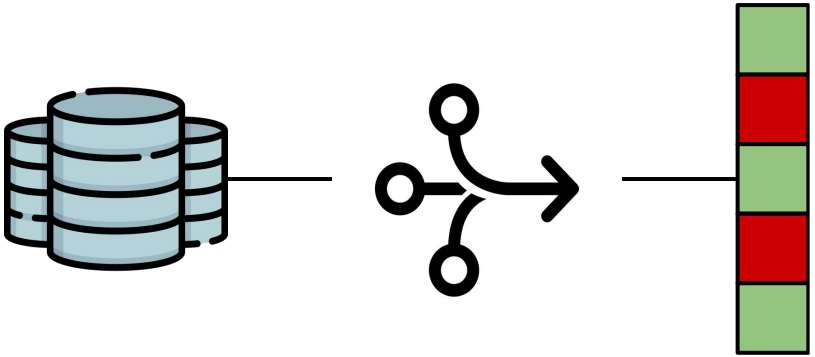


Unified Assessment

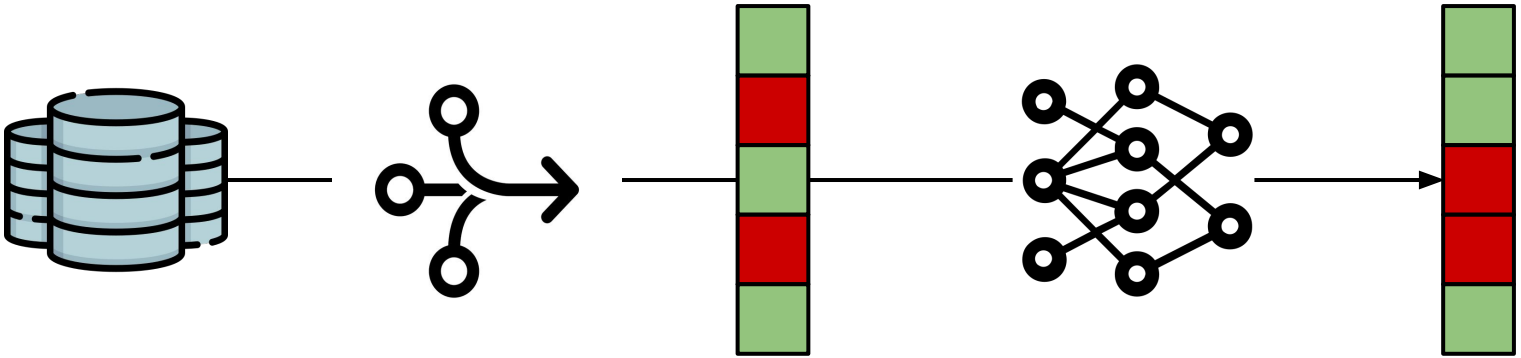
Snippets



CODE READABILITY



CODE READABILITY



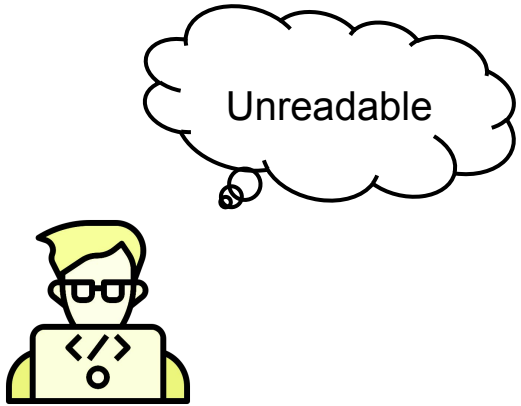
CODE READABILITY

Readability is indeed subjective

CODE READABILITY

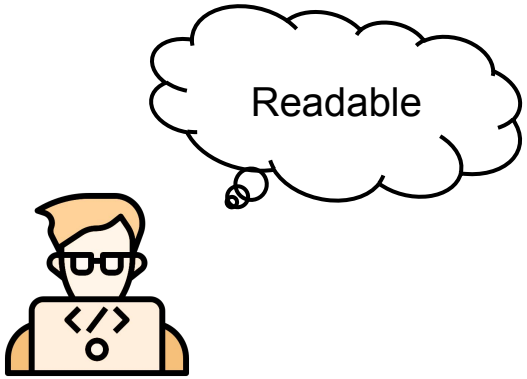
```
xsp = jj_scanpos;
if (jj_scan_token (100) ) {
jj_scanpos = xsp;
if (jj_scan_token (101) ) return true;
```

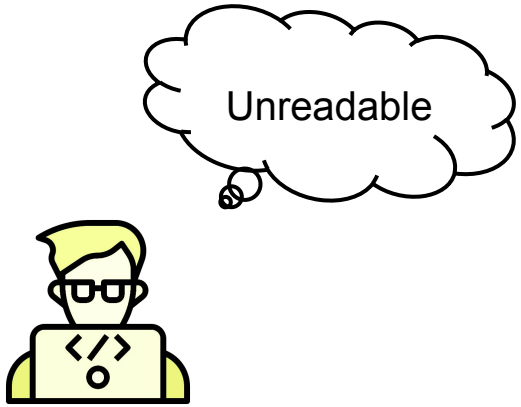
CODE READABILITY




```
xsp = jj_scanpos;
if (jj_scan_token (100) ) {
jj_scanpos = xsp;
if (jj_scan_token (101) ) return true;
```

CODE READABILITY






```
xsp = jj_scanpos;
if (jj_scan_token (100) ) {
jj_scanpos = xsp;
if (jj_scan_token (101) ) return true;
```

CODE READABILITY

To what extent is it possible to **assess code readability** as **subjectively** perceived by developers?

CODE READABILITY

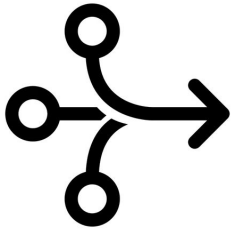
Developers Assessments

Snippets					
	3	3	1	4	5
	3	1	3	3	4
	2	3	3	2	3
	1	3	1	3	2
	4	4	5	5	4

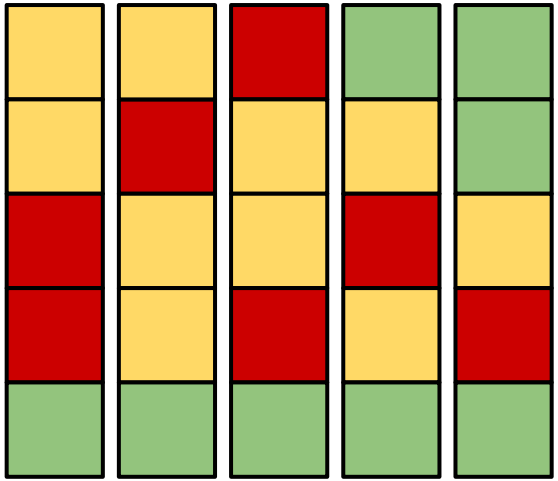
CODE READABILITY

Developers Assessments

Snippets					
	3	3	1	4	5
	3	1	3	3	4
	2	3	3	2	3
	1	3	1	3	2
	4	4	5	5	4



Developers Assessments



Unreadable - {1, 2}

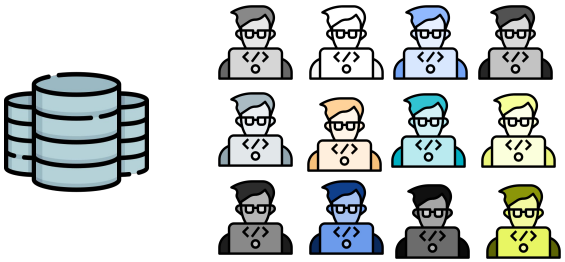
Neutral - {3}

Readable - {4, 5}

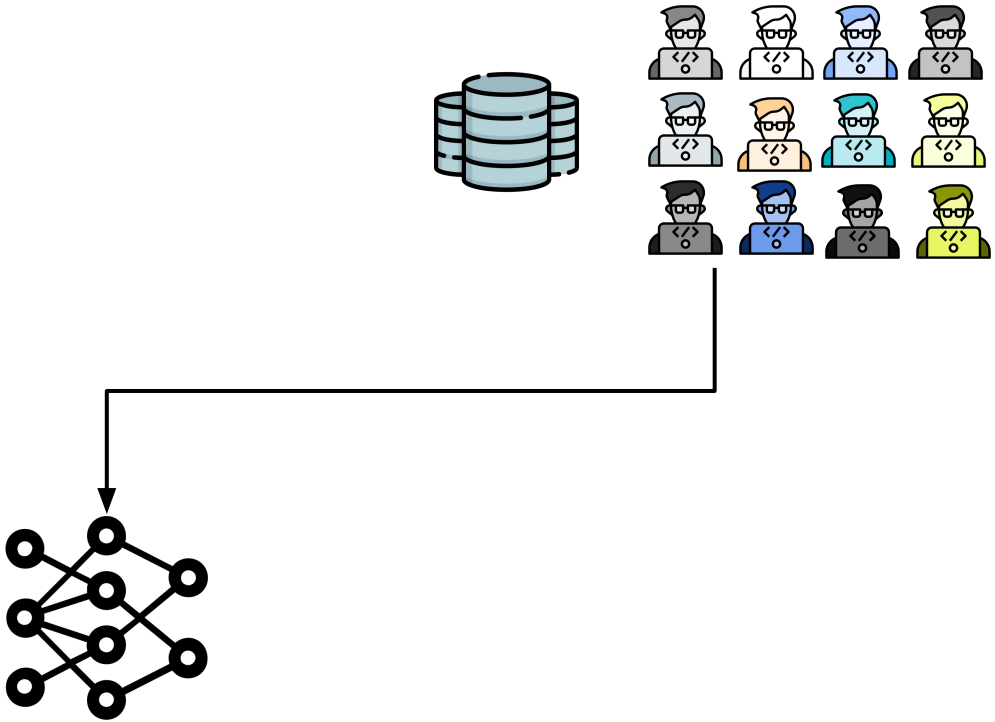
RQ1

How effective are **generalist models** in predicting developers' subjective assessments of code readability

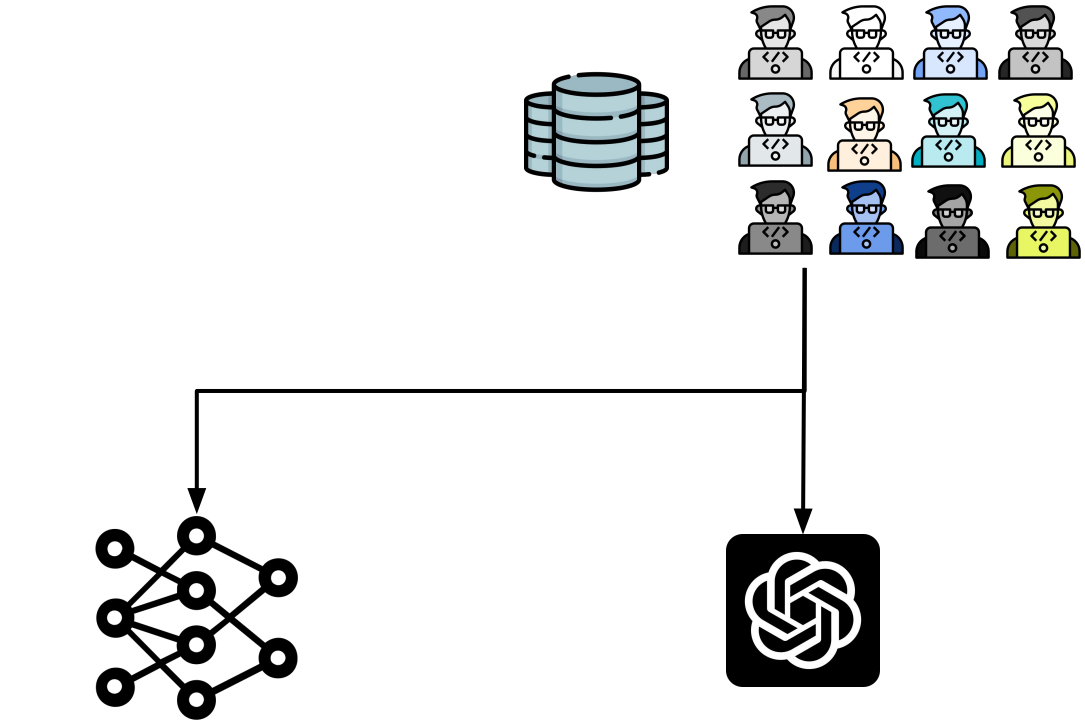
RQ1



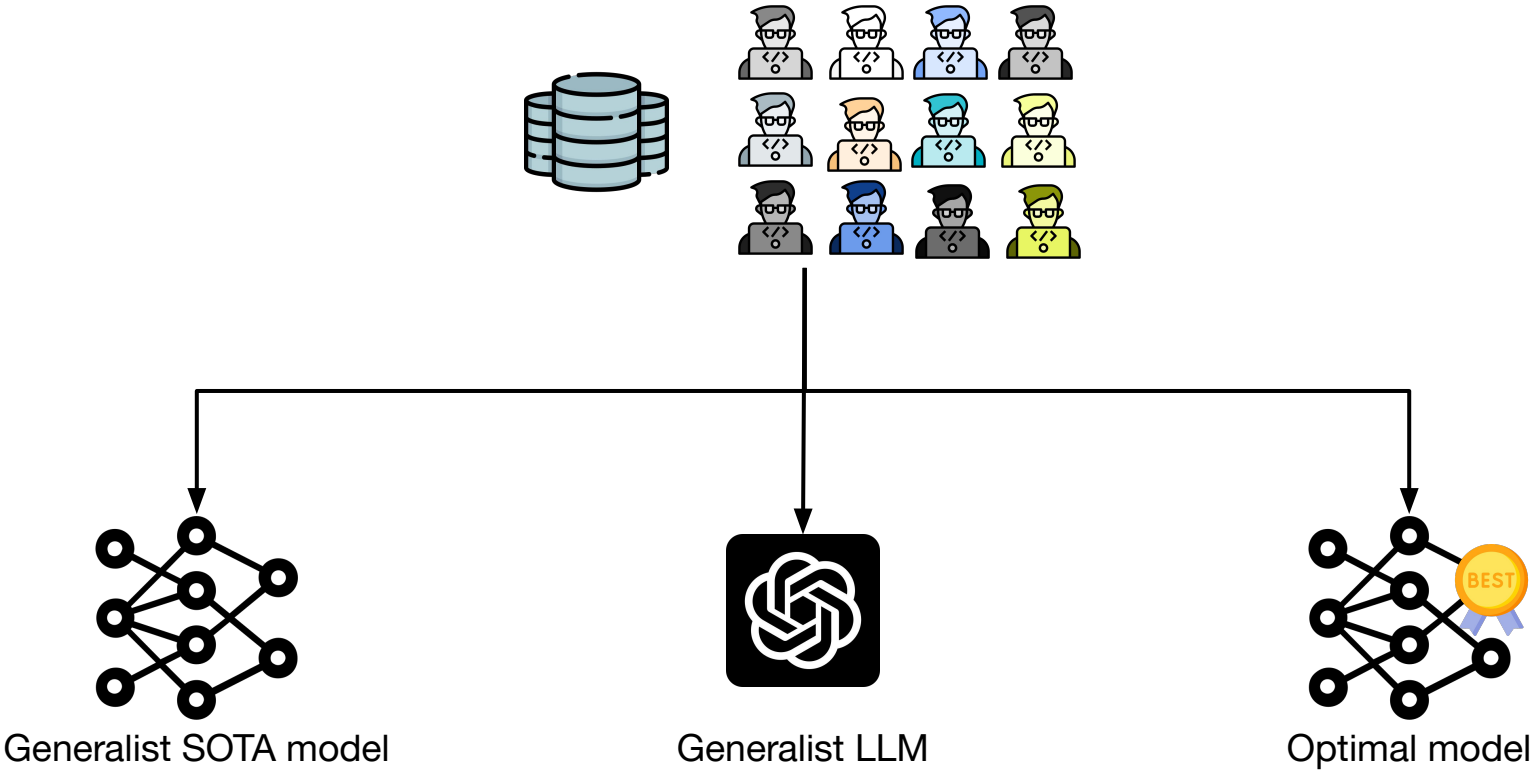
RQ1



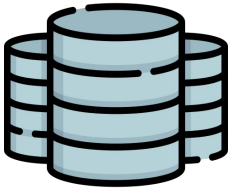
RQ1



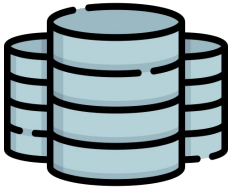
RQ1



RQ1

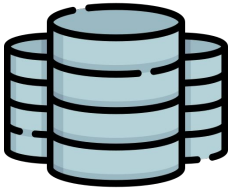


D_s

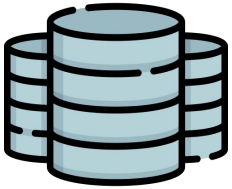


$D_{b\&w}$

RQ1

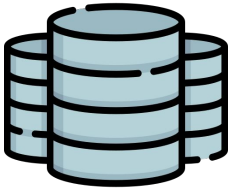


D_s

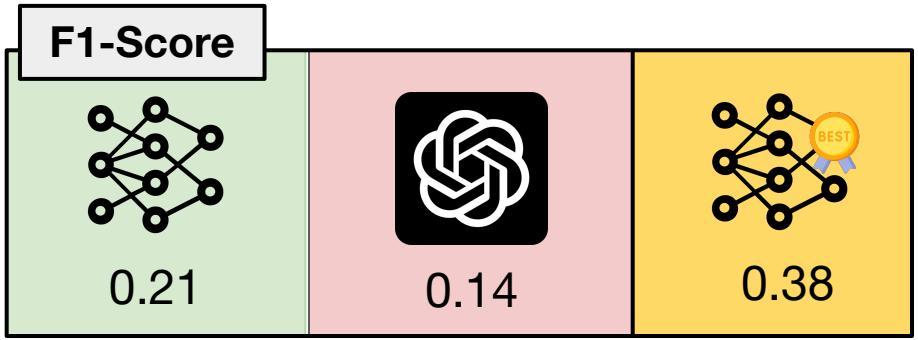


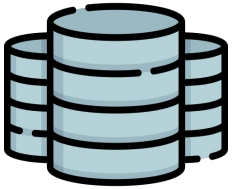
$D_{b\&w}$

RQ1

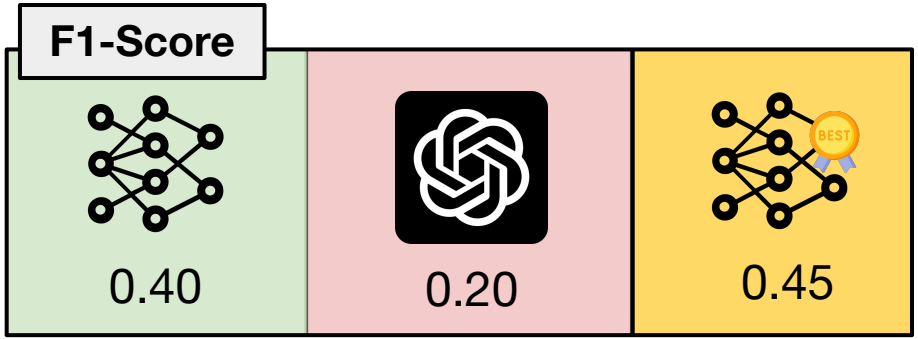


D_s





$D_{b\&w}$



RQ1

Generalist models are not effective for personalized code readability assessments.

RQ1

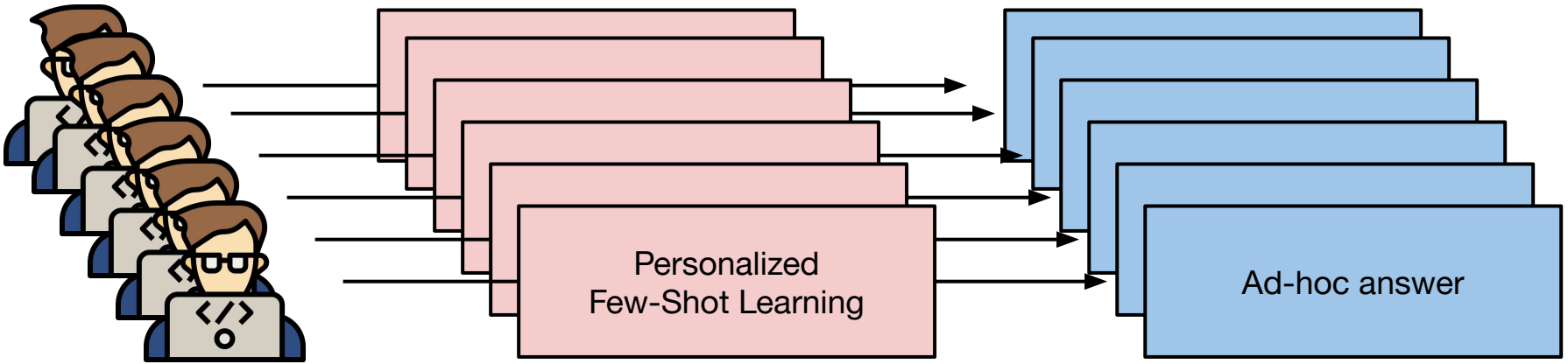
Generalist models are not effective for personalized code readability assessments.

Generalist LLM achieves lower performance than a significantly smaller, feature-based model tailored for readability predictions.

RQ2

How **effective** is an LLM-based **personalized** code readability assessment model in predicting developers subjective assessments of code readability?

RQ2



RQ2

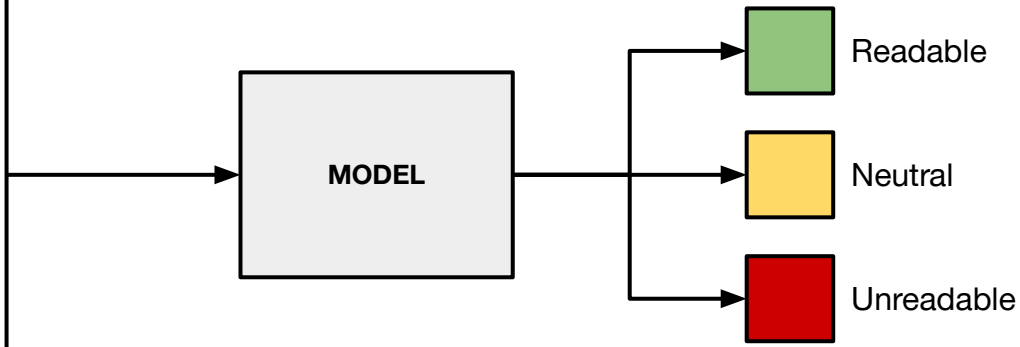
MODEL INPUT

You are an expert and personal code readability labeler.

Your role is to assign a code readability label based on the already known preferences of the developer. The labels you can assign are: *Unreadable*, *Neutral*, and *Readable*.

Below you find examples of the already known developer preferences.
`0-shot`
`1-shot`
`2-shot`

Now, assign the developer code readability label for the following one:
`target-snippet`



RQ2

MODEL INPUT

You are an expert and personal code readability labeler.

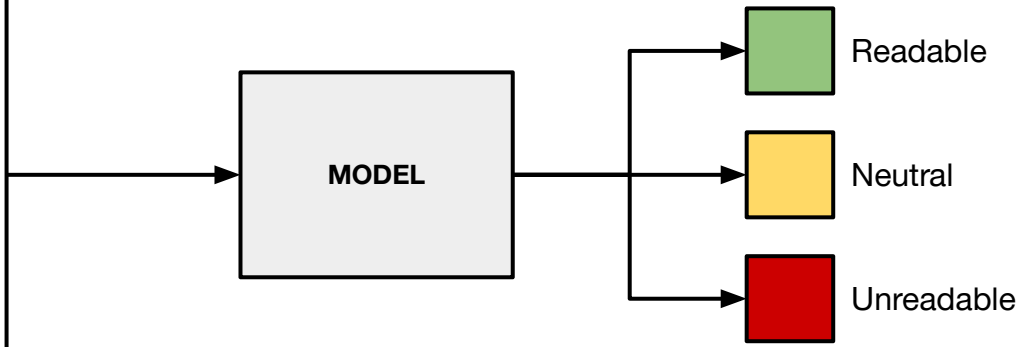
Your role is to assign a code readability label based on the already known preferences of the developer. The labels you can assign are: *Unreadable*, *Neutral*, and *Readable*.

Below you find examples of the already known developer preferences.

```
0-shot
1-shot
2-shot
```

Now, assign the developer code readability label for the following one:

```
target-snippet
```



RQ2

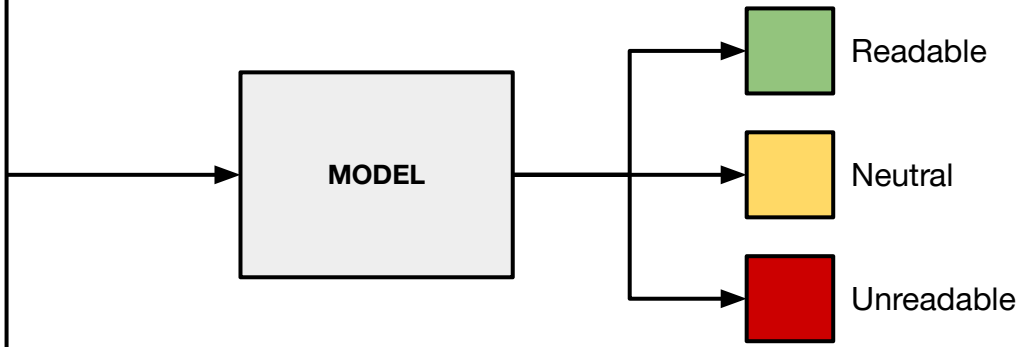
MODEL INPUT

You are an expert and personal code readability labeler.

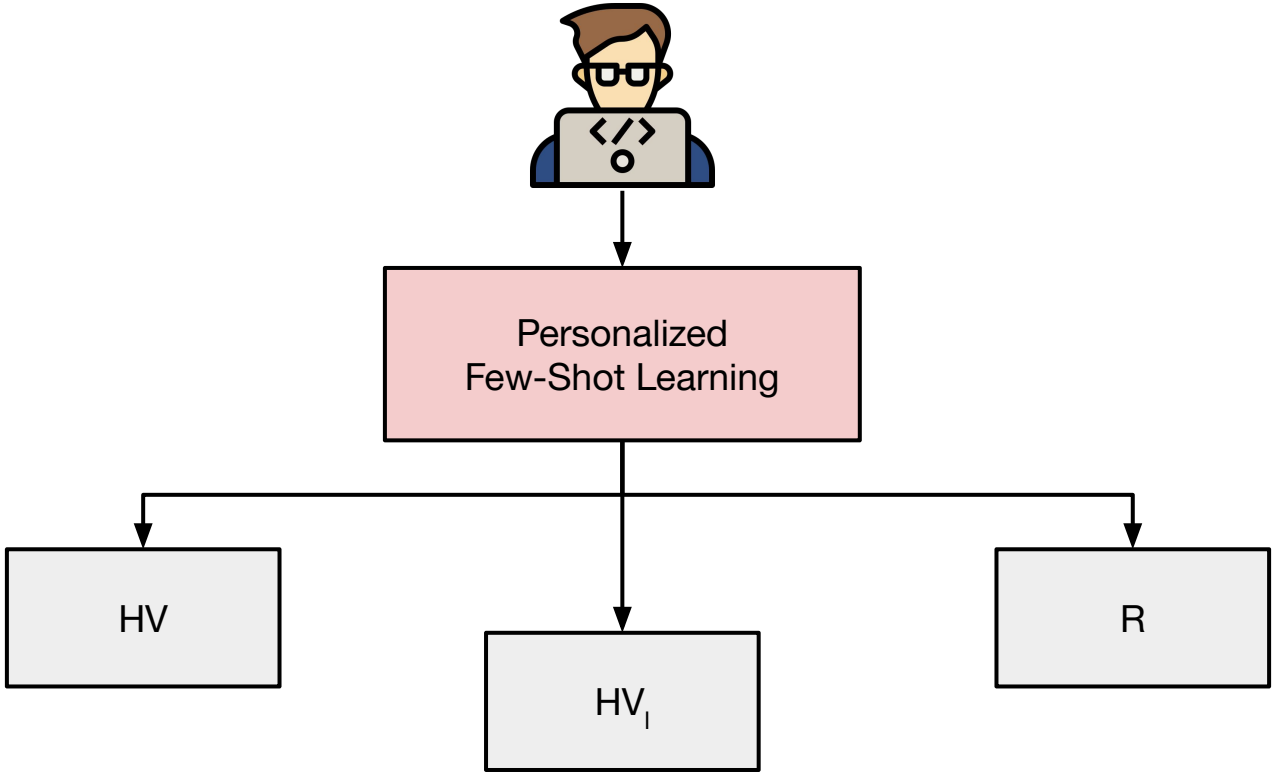
Your role is to assign a code readability label based on the already known preferences of the developer. The labels you can assign are: *Unreadable*, *Neutral*, and *Readable*.

Below you find examples of the already known developer preferences.
`0-shot`
`1-shot`
`2-shot`

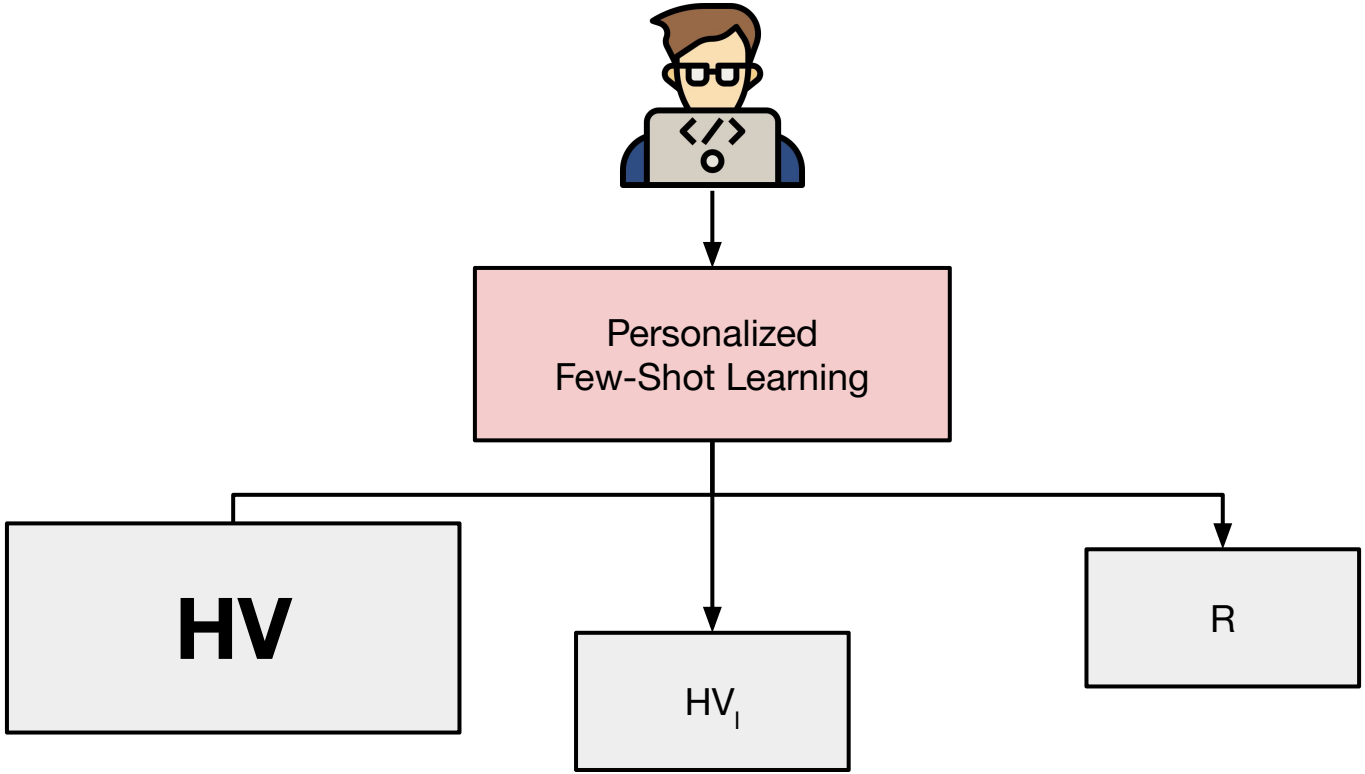
Now, assign the developer code readability label for the following one:
`target-snippet`



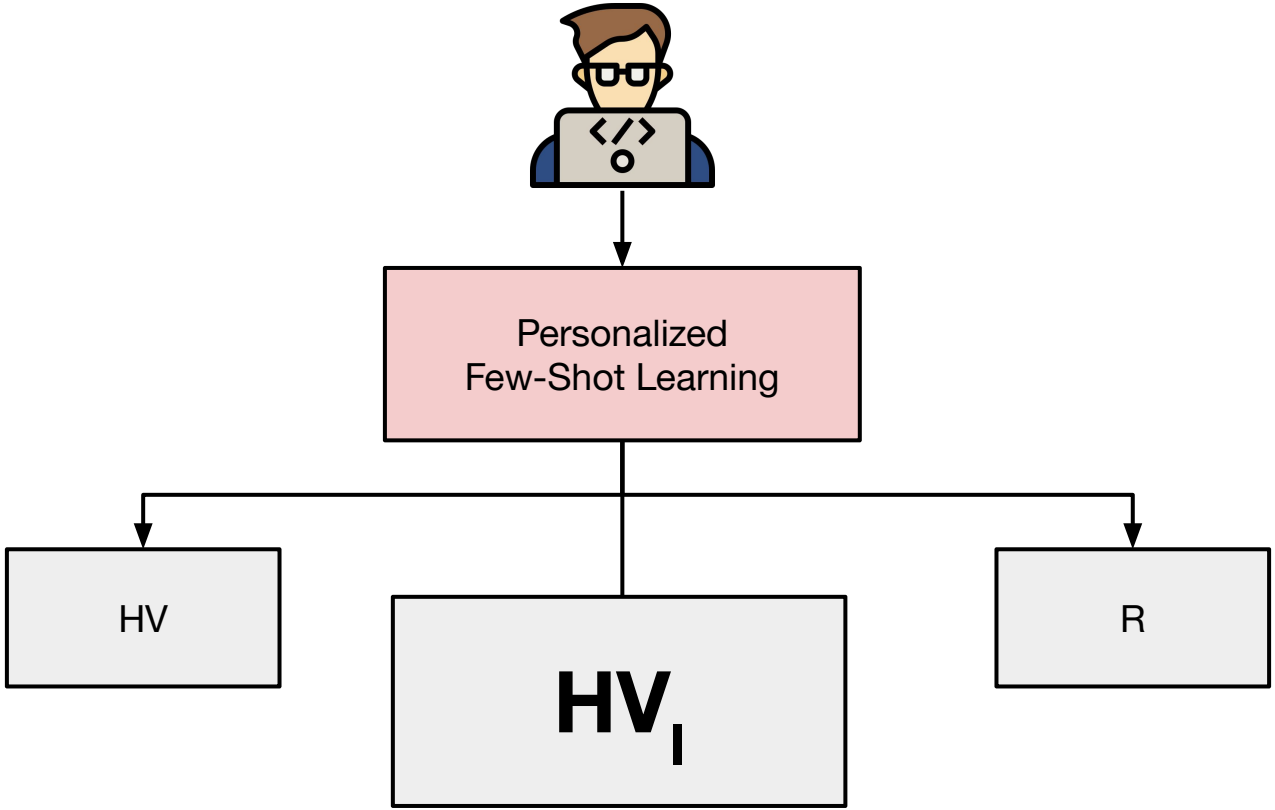
RQ2



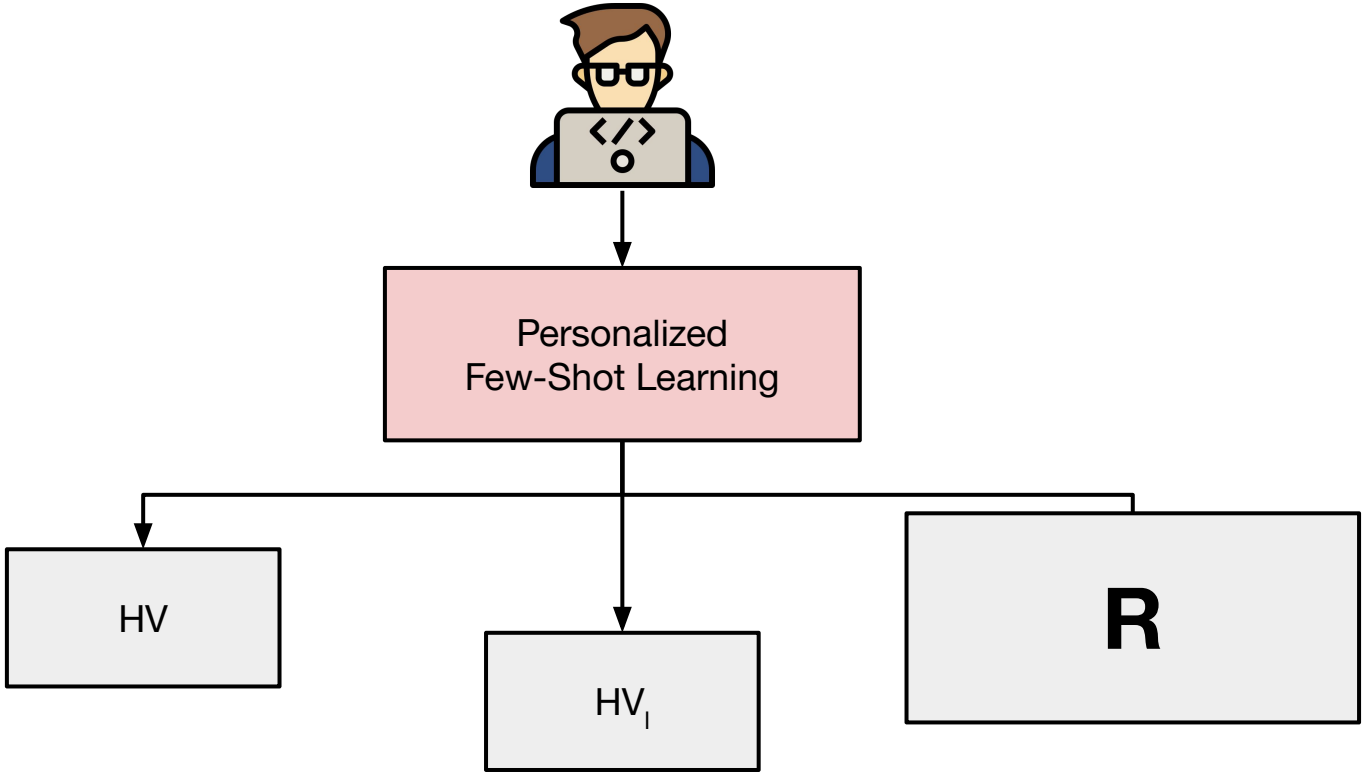
RQ2



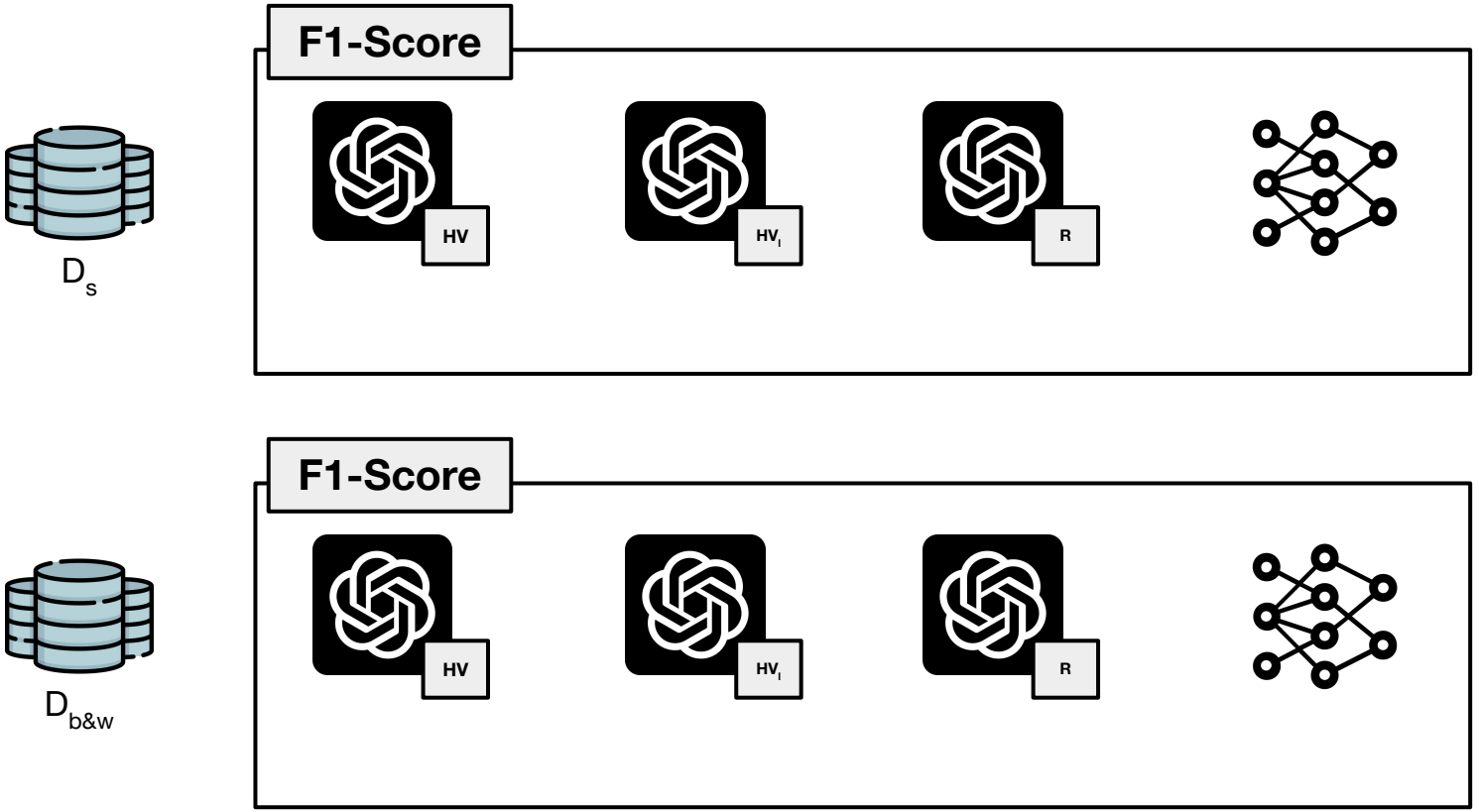
RQ2



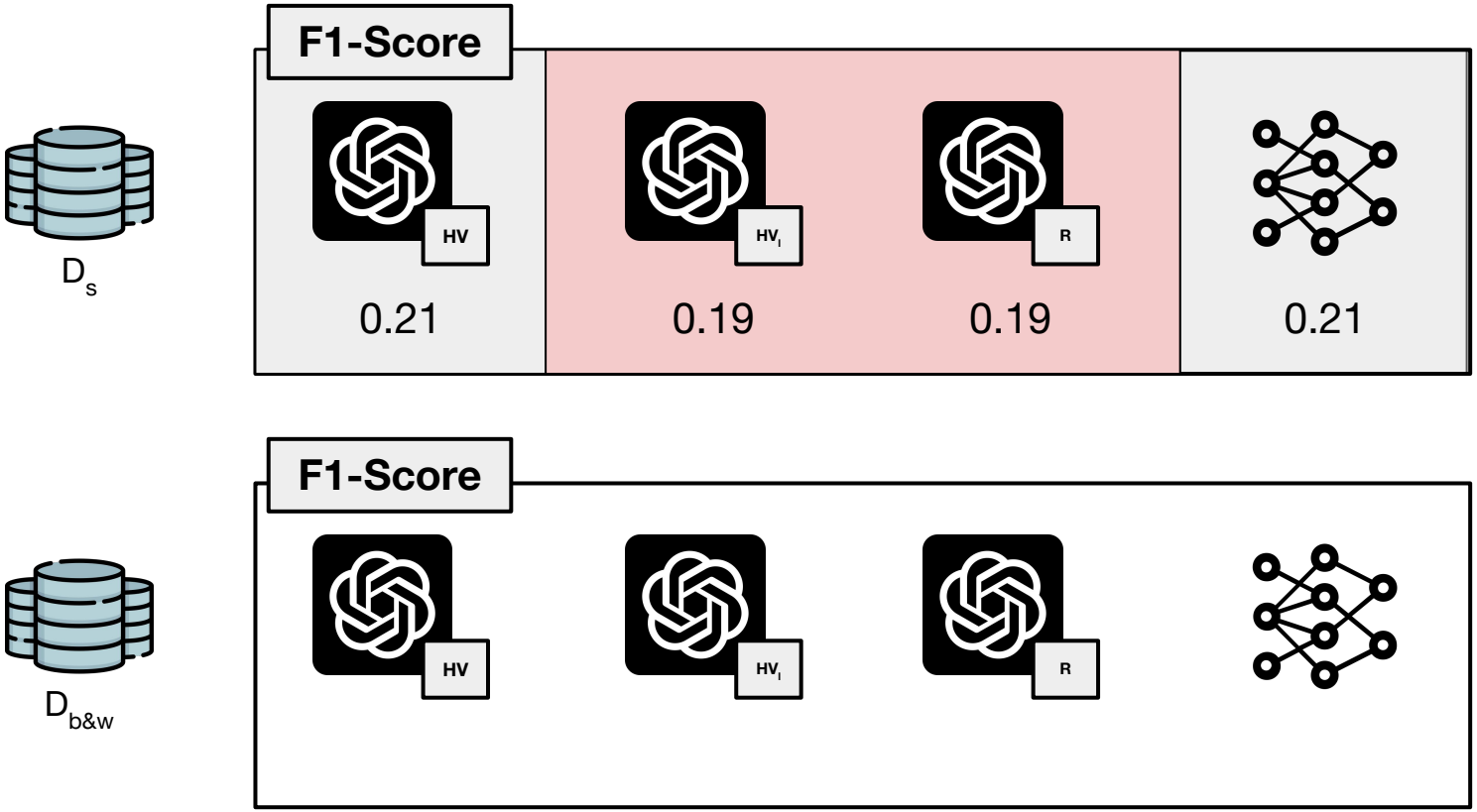
RQ2



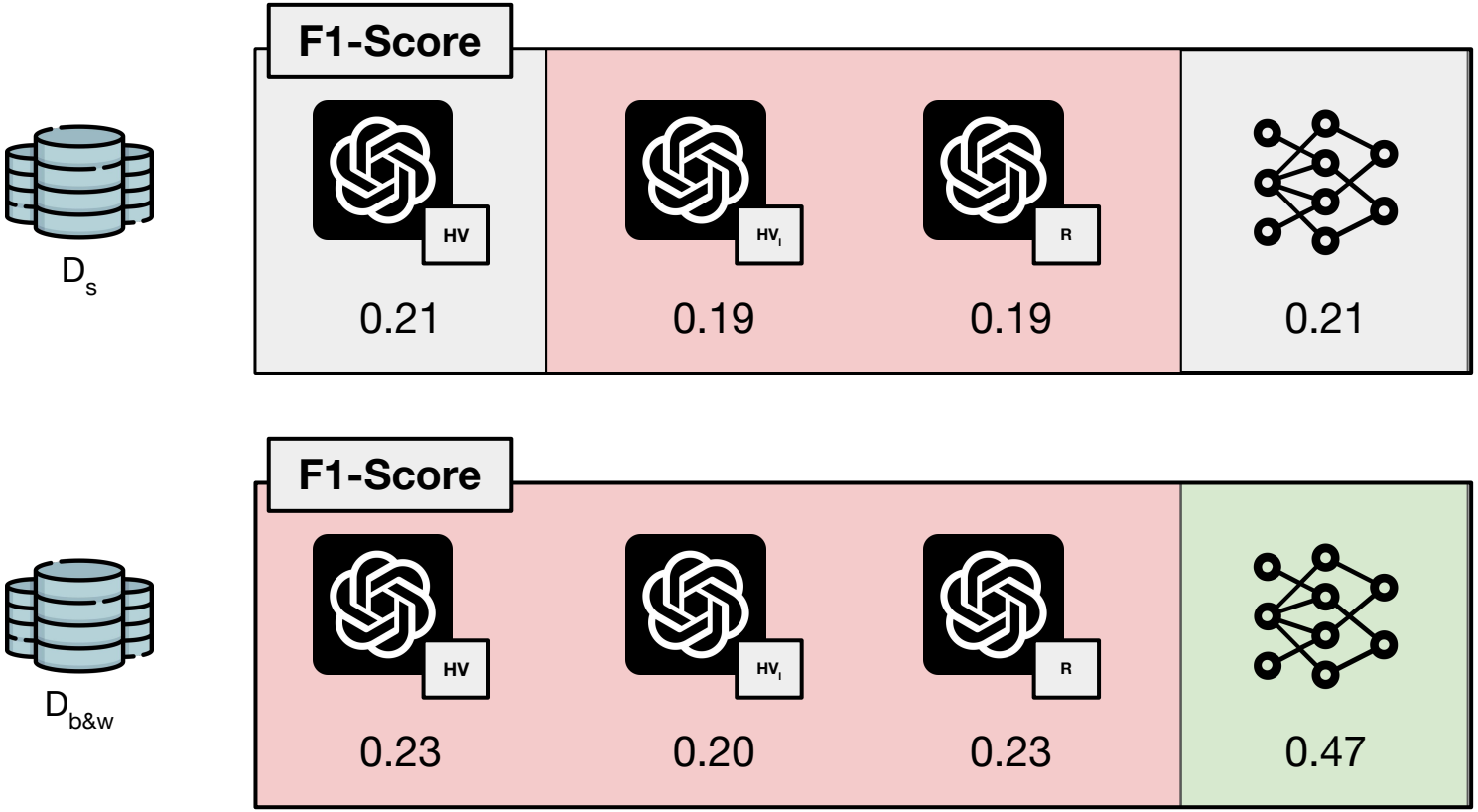
RQ2



RQ2



RQ2


F1-Score
HV
HV$_I$
R
0.21
0.19
0.19
0.21
D_s
F1-Score
HV
HV$_I$
R
0.23
0.20
0.23
0.47
$D_{b\&w}$

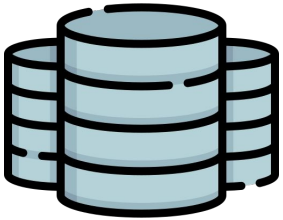
RQ2

Personalized LLM-based code readability models are **less effective** than a state-of-the-art generalist model.

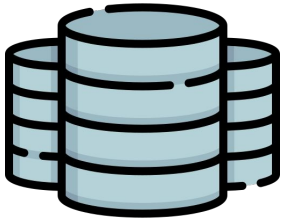
RQ3

To what extent are developers' code readability assessments **consistent**

RQ3



$D_{b\&w}$



D_s

384 samples

384 samples

RQ3

Developers code readability assessments show moderate consistency, with **32% inconsistency in $D_{b\&w}$ and 23% in D_s** .

OUTCOMES

The definition of **personalized code readability prediction models** is worth future investigation.

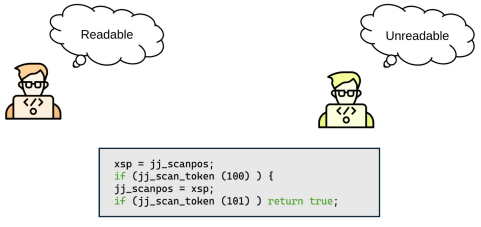
OUTCOMES

The definition of **personalized code readability prediction models** is worth future investigation.

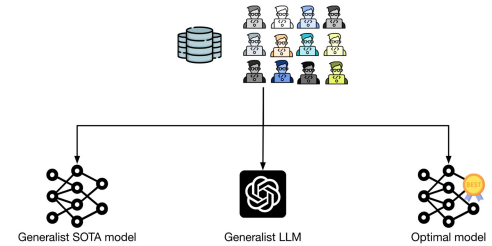
Future work should aim at using more principled procedures for defining code readability datasets.

SUMMARY

CODE READABILITY



RQ1

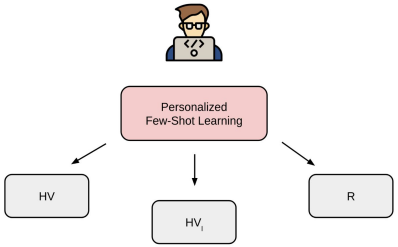


RQ1

Generalist models are not effective for personalized code readability assessments.

Generalist LLM achieves lower performance than a significantly smaller, feature-based model tailored for readability predictions.

RQ2



RQ2

Personalized LLM-based code readability models are **less effective** than a state-of-the-art generalist model.

RQ3

Developers code readability assessments show moderate consistency, with 32% **inconsistency in $D_{b\&w}$ and 23% in D_s** .