

Code Review Automation: Strengths and Weaknesses of the State of the Art

Rosalia Tufano , *Student Member, IEEE*, Ozren Dabić , Antonio Mastropaolo , *Student Member, IEEE*, Matteo Ciniselli , *Graduate Student Member, IEEE*, and Gabriele Bavota , *Member, IEEE*

Abstract—The automation of code review has been tackled by several researchers with the goal of reducing its cost. The adoption of deep learning in software engineering pushed the automation to new boundaries, with techniques *imitating* developers in generative tasks, such as commenting on a code change as a reviewer would do or addressing a reviewer’s comment by modifying code. The performance of these techniques is usually assessed through quantitative metrics, *e.g.*, the percentage of instances in the test set for which correct predictions are generated, leaving many open questions on the techniques’ capabilities. For example, knowing that an approach is able to correctly address a reviewer’s comment in 10% of cases is of little value without knowing what was asked by the reviewer: What if in all successful cases the code change required to address the comment was just the removal of an empty line? In this paper we aim at characterizing the cases in which three code review automation techniques tend to succeed or fail in the two above-described tasks. The study has a strong qualitative focus, with ~ 105 man-hours of manual inspection invested in manually analyzing correct and wrong predictions generated by the three techniques, for a total of 2,291 inspected predictions. The output of this analysis are two taxonomies reporting, for each of the two tasks, the types of code changes on which the experimented techniques tend to succeed or to fail, pointing to areas for future work. A result of our manual analysis was also the identification of several issues in the datasets used to train and test the experimented techniques. Finally, we assess the importance of researching in techniques specialized for code review automation by comparing their performance with ChatGPT, a general purpose large language model, finding that ChatGPT struggles in commenting code as a human reviewer would do.

Index Terms—Automated code review, empirical study.

I. INTRODUCTION

THE benefits of code review are well-known and supported by empirical evidence [8], [9], [27], [29], [39]. However, works in the literature also documented the non-trivial costs

of such a process [11], [36], [37]. To bring down such a cost, researchers proposed techniques to (partially) automate code review tasks. Originally, most of the approaches aimed at recommending appropriate reviewers for a given code change [5], [7], [19], [20], [28], [30], [35], [43], [52], [56], [57]. With the rise of Deep Learning (DL) in software engineering, researchers started addressing more challenging generative tasks aimed at *imitating* human developers. Two concrete examples are the tasks automated by Tufano et al. [48]¹ using a Transformer model. The first, named *code & comment-to-code*, aims at automatically implementing a revised code (C_r) given the code submitted for review (C_s) and a reviewer comment (R_{nl}). Thus, the input of the approach is represented by a pair $\langle C_s, R_{nl} \rangle$, while the output is C_r .

The second, named *code-to-comment*, aims at commenting a code submitted for review as a reviewer would do: The input of the approach is C_s , while the expected output is R_{nl} (*i.e.*, a natural language comment asking for changes to C_s).

The work by Tufano et al. [48] is only the first of several focusing on automating one or both of these tasks [18], [22], [23], [48]. These techniques have been evaluated on test sets containing hundreds of instances representative of the automated tasks. For example, for the *code & comment-to-code* task, the test sets feature $\langle C_s, R_{nl} \rangle$ pairs which are fed to the approach to assess whether it can address the reviewer’s comment R_{nl} and generate the expected C_r . The outcome of these evaluations is a mostly quantitative report showing, *e.g.*, the percentage of instances in the test set for which the approach successfully generated a prediction. However, such quantitative measures only tell part of the story. Indeed, it could happen that the approach is targeting the *low-hanging fruits*, being successful in only simple code review scenarios which are unlikely to save developers’ time. For the *code & comment-to-code* task this might mean successfully addressing mostly comments requiring minor changes to C_s (*e.g.*, addition/removal of whitespaces to improve the formatting). Similarly, for the *code-to-comment* task the approach could overfit and mostly be successful in posting comments related to *e.g.*, replacing the `==` operator in Java with an `equals` invocation when needed. In other words, little is known about the code review scenarios in which these techniques succeed or fail.

¹We refer to our previous works in the area as Tufano et al. because the set of authors only partially overlaps.

Manuscript received 2 June 2023; revised 8 November 2023; accepted 25 December 2023. Date of publication 1 January 2024; date of current version 13 February 2024. This work was supported by the European Research Council (ERC) through the European Union’s Horizon 2020 research and innovation programme under Agreement 851720. Recommended for acceptance by R. Hoda. (Corresponding author: Gabriele Bavota.)

The authors are with SEART Software Institute, Università della Svizzera Italiana, 6900 Lugano, Switzerland (e-mail: rosalia.tufano@usi.ch; ozren.dabic@usi.ch; antonio.mastropaolo@usi.ch; matteo.ciniselli@usi.ch; gabriele.bavota@usi.ch).

Digital Object Identifier 10.1109/TSE.2023.3348172

TABLE I
SUMMARY OF RELATED WORK. FOR LI Z. ET AL. [23] THE SIZE OF THE TEST SET REFERS TO JAVA INSTANCES ONLY

Reference	Venue	Task	Granularity	Underlying technique	# Training instances	# Test instances
Tufano M. et al. [46]	ICSE'19	<i>code-to-code</i>	method	NMT	8.6k	1k
Tufano R. et al. [49]	ICSE'21	<i>code-to-code</i> <i>code & comment-to-code</i>	method	NMT	13.7k	1.7k
Tufano R. et al. [48]	ICSE'22	<i>code-to-code</i> <i>code & comment-to-code</i> <i>code-to-comment</i>	method	T5 (pre-trained)	134.2k	16.7k
Thongtanunam et al. [42]	ICSE'22	<i>code-to-code</i>	method	NMT	118k	14.7k
Li L. et al. [22]	ESEC/FSE'22	<i>code & comment-to-code</i> <i>code-to-comment</i>	method	T5 (pre-trained)	87k	1k
Hong et al. [18]	ESEC/FSE'22	<i>code-to-comment</i>	method	IR	13.7k	1.7k
Li Z. et al. [23]	ESEC/FSE'22	<i>code & comment-to-code</i> <i>code-to-comment</i> <i>code quality estimation</i>	diff hunk	T5 (pre-trained)	117.7k 150.4k 265.8k	2.2k 1.6k 66.4k

To fill this gap, we manually analyzed 2,291 predictions generated by three state-of-the-art techniques [18], [23], [48] automating the code review tasks previously described. The predictions have been generated on the original test sets used in the papers presenting the subject techniques.

The result of such an analysis are two taxonomies (one per task) featuring a total of 120 types of code changes requested during code review (e.g., *extract method refactoring*, *add thrown exception*) with indication about the extent to which state-of-the-art techniques are successful in (i) requesting their implementation when needed by properly commenting the submitted code as a human reviewer would do (*code-to-comment* task); (ii) automatically implementing them to address a reviewer's comment (*code & comment-to-code* task).

We found that the proposed techniques can provide support in a wide variety of code changes. However, there are areas of our taxonomies in which the approaches consistently fail, pointing to the need for more research. As a concrete example, the experimented techniques struggle when they need to recommend (*code-to-comment* task) or implement (*code & comment-to-code* task) complex code changes spanning across several code components. This is due to the “view” they have of the code base, usually limited to a single function or diff hunk submitted for review. This indicates the need for enriching the contextual information provided to these techniques.

During our manual analysis we also found that ~25% of the instances in the inspected datasets are the result of data extraction errors possibly undermining the techniques' performance and questioning the validity of the evaluation performed on the test set. We discuss the reasons for such problematic instances.

Finally, given the recent proposal of general purpose Large Language Models (LLMs) such as ChatGPT [1], it is unclear what the actual need is for code review automation via specialized techniques. We compared the three subject approaches with ChatGPT, showing that, while the latter represents a competitive solution for the *code & comment-to-code* task, it suffers in the *code-to-comment* task.

We release [4] all data used in and output of our study.

II. RELATED WORK

Most of the works in code review automation aim at recommending the most suited reviewer for a given change [5], [7], [12], [19], [20], [28], [30], [35], [41], [43], [52], [53], [54], [56], [57]. These works differ in the features and algorithms used for the recommendation. Other techniques focus on the binary classification of the quality of the code submitted for review (i.e., whether the change should be accepted or not) using machine [12] or deep [23], [40] learning. Our work is however mostly related to the techniques aimed at *imitating* human reviewers by automatically reviewing the submitted code. Thus, we focus our discussion on these techniques which are summarized in Table I.

Tufano M. et al. [46] proposed the usage of Neural Machine Translation (NMT) to learn how to automatically modify a given Java method as developers would do during a pull request (PR): The NMT model takes as input a method submitted in a PR and implements code changes likely to be required during the review of the PR.

Tufano R. et al. [49] built on top of this idea by presenting two transformer-based models to automate two code review tasks. The first, named *code-to-code*, is a replica of the task automated by Tufano M. et al. [46]: It takes as input a Java method submitted for review (C_s) and implements changes likely to be required during code review, producing a revised method C_r . The second is the already discussed *code & comment-to-code* task.

Both of the aforementioned works suffered of a major limitation: They relied on a code abstraction process to reduce the vocabulary size and simplify the learning of the DL models. This means that the models did not work on raw code, but on an abstracted version of it in which, for example, variable names were replaced with a VAR_ID token, with ID being a progressive number going from 1 to n (n = number of variables declared in the method). However, such an abstraction had a cost to pay, since the models were not able to implement code changes requiring the introduction of new identifiers and

literals. To understand this point let us consider the *code-to-code* task, in which the model takes as input a submitted method C_s and generates a revised method C_r . Both the input and the generated method are abstracted. The abstraction process applied to C_s allows to map each identifier to its abstracted version. For example, it is possible to know that a variable `sum` in C_s has been mapped to `VAR_1`. Thus, if the generated C_r method features `VAR_1`, it can be mapped back to `sum`, making C_r understandable by the developer receiving such a recommendation. Differently, if C_r features a new identifier `VAR_4` which was not present in C_s , it is not possible to map it to raw code, making the recommendation useless. For this reason, both of these works [46], [49] only automated very simple code review tasks, mostly requiring re-arranging the code tokens in C_s and avoiding the introduction of new identifiers and literals.

To address this limitation, Tufano R. et al. [48] proposed a new approach which exploited the Text-To-Text Transfer Transformer (T5) model [34] and adopted a SentencePiece tokenizer [21], allowing the model to work with raw source code by keeping under control the vocabulary size, avoiding the need for abstraction. Besides reporting better performance as compared to their previous approach [49], this work is also the first one attempting the automation of the *code-to-comment* task (*i.e.*, commenting the code to request changes as a human reviewer would do).

Thongtanunam et al. [42] also used transformers to automate code review. They focused on the *code-to-code* task, comparing with the original work by Tufano M. et al. [46] and showing the superiority of their approach.

Li L. et al. [22] relied on pre-trained DL models to improve the performance achieved by Tufano R. et al. [48] on the *code & comment-to-code* and the *code-to-comment* tasks. Hong et al. [18] showed instead that a simpler approach based on IR can achieve competitive performance as compared to DL models when it comes to the *code-to-comment* task: Rather than generating a reviewer's comment for a submitted Java method (as done by the DL-based techniques), the approach by Hong et al. [18] identifies in the "training set" the method C_{ms} being most similar to the C_s submitted for review. Based on this information, it retrieves review comments which have been posted in the past for C_{ms} , assuming they would be valuable for C_s as well. Both Li L. et al. [22] and Hong et al. [18] show the superiority of their technique as compared to the approach by Tufano R. et al. [48].

Finally, Li Z. et al. [23] targeted the automation of three code review tasks. The first among these represent the binary classification of submitted code to decide whether it needs a review or can be accepted as is. The other two tasks are *code-to-comment* and *code & comment-to-code*. One aspect makes this work particularly novel as compared to those previously discussed: While all previous techniques work at method-level granularity (*i.e.*, they take as input a method submitted for review), the approach by Li Z. et al. [23] takes a "diff hunk", namely an area of a specific file modified in a code change to be reviewed. This allows the model to generate file-level "reviewers' comments" such as those asking changes to the

`import` statements, which are ignored by previous techniques. Given also its ability to work with code written in nine different programming languages, this is currently considered the state-of-the-art technique.

III. STUDY DESIGN

The *goal* of this study is to assess the capabilities of state-of-the-art techniques for code review automation. The *context* consists of: (i) three techniques, *i.e.*, Tufano R. et al. [48] (T5CR), Li Z. et al. [23] (CODEREVIEWER), and Hong et al. [18] (COMMENTFINDER); (ii) two code review tasks for which the subject techniques provide automation, *i.e.*, *code-to-comment* and *code & comment-to-code*; and (iii) instances featured in the test datasets on which the techniques have been evaluated in the papers presenting them. We do not aim at comparing the performance of the three techniques to understand which one is the best. Rather, we look at them as a whole to understand the status of code review automation.

We address the following research questions (RQs):

RQ1: What are the characteristics of correct and wrong recommendations generated by techniques for code review automation? We cluster the predictions generated by the experimented techniques into two sets representing instances for which the approaches generated a correct or a wrong prediction. Then, we quantitatively compare these two sets. For the *code-to-comment* task we compare the "complexity" of the comments to automatically generate (*i.e.*, those present in the ground truth). Similarly, for the *code & comment-to-code* task, we compare the "complexity" of the code changes to implement. Such an analysis will shed some light on the extent to which the state-of-the-art techniques overfit towards the low-hanging fruits of the datasets.

On top of this, we qualitatively analyze 2,291 predictions generated by the three approaches (equally distributed between correct and wrong predictions) to characterize the type of code change they were able to request (*code-to-comment* task) or to automatically implement (*code & comment-to-code* task). The objective is to understand the scenarios in which these techniques are successful *vs* those in which they tend to fail. For example, if the outcome reveals that for the *code & comment-to-code* task the techniques are mostly successful in implementing formatting changes, but tend to fail when dealing with more challenging code changes (*e.g.*, fixing a bug), this would question their usefulness.

RQ2: To what extent are the datasets used to train and test techniques for code review automation suitable for such a scope? Through qualitative analysis we unveil the presence of problematic instances in the datasets used in the subject studies, calling for better dataset-cleaning pipelines.

RQ3: How do techniques for code review automation proposed in the literature compare to state-of-the-art large language models? We compare the three subject techniques with ChatGPT [1] as representative of LLMs. Such a comparison informs the need to further invest in automating code review through specialized models rather than relying on general-purpose LLMs.

TABLE II
INSPECTED INSTANCES

Reference	Task	# Correct (%)	# Wrong (%)	Inspected Correct	Inspected Wrong	Valid Correct	Valid Wrong
T5CR [48]	<i>code&comment-to-code</i>	2'363 (14.08%)	14'417 (85.92%)	178	272	199	167
	<i>code-to-comment</i>	354 (2.11%)	16'426 (97.89%)	200	227	169	189
COMMENTFINDER [18]	<i>code-to-comment</i>	479 (2.85%)	16'301 (97.15%)	234	254	169	176
CODEREVIEWER [23]	<i>code&comment-to-code</i>	599 (27.15%)	1'607 (72.85%)	197	317	197	176
	<i>code-to-comment</i>	0 (0%)	1'611 (100%)	-	412	50	179

A. Study Context

We present the study context in terms of experimented techniques and datasets/predictions we analyzed.

1) *Techniques for Code Review Automation*: Our work focuses on techniques aimed at *imitating* human reviewers, thus automating the *code-to-comment* and/or *code & comment-to-code* task. Based on our analysis of the literature, five approaches target these tasks: Tufano R. et al. [49], T5CR [48], COMMENTFINDER [18], CODEREVIEWER [23], and Li L. et al. [22]. Tufano R. et al. [49] has been excluded since in their followup work [48] the authors showed the superiority of the newly presented technique (T5CR) as compared to their first attempt [49] in automating code review. Li L. et al. [22], instead, has been excluded after inspecting the replication package shared by the authors: We rely on the authors' replication packages to download (and then inspect) the predictions generated by their model on the test set. The replication package for Li L. et al. [22] featured 987 predictions for the 1,055 instances in the test set, casting doubts on the mapping between test instances and predictions. Also, 7 of these predictions had one or more "partial duplicate" in the training set, meaning that the training set featured the same code instance (*i.e.*, the same code for which the technique had to automatically generate "reviewer's comments") of an entry in the test set with a different target message. While this is not a problem in principle, this plays a role in our qualitative evaluation, where we analyze whether the comments generated by these techniques are semantically equivalent to the expected one (despite they might use a different wording). Having the same code instance in the training set allows the approach by Li L. et al. [22] to "reuse" the reviewer's comment from the training set, thus generating something that, for sure, will be meaningful. For these reasons we discarded this technique from our study. This left us with the three techniques.

For the *code-to-comment* task, we consider predictions generated by all three approaches with: T5CR [48] being representative of DL-based techniques working at method-level granularity and not considering the code diff as an input; COMMENTFINDER [18] being representative of IR-based techniques also working at method-level granularity; and CODEREVIEWER [23] being representative of DL-based techniques working at "diff hunk" granularity and considering the code diff as an input. For the *code & comment-to-code* task we only consider T5CR and CODEREVIEWER, since COMMENTFINDER does not provide support for such a task.

2) *Datasets and Predictions*: From the replication packages of the three techniques [18], [23], [48] we collected the test sets used for their evaluation and the corresponding predictions. The

size of the test datasets is reported in Table I. There are two main differences among the datasets. The first is in the representation of the code submitted for review that is provided as input to the technique (C_s). While for T5CR [48] and COMMENTFINDER [18] C_s is a Java method, for CODEREVIEWER [23] is a diff hunk. The second concerns the fact that T5CR and COMMENTFINDER only work with Java code, while CODEREVIEWER supports nine languages, including Java. In our study we only considered Java instances for consistency and to simplify the following described manual analysis.

B. Data Collection and Analysis

1) RQ_1 : *Correct vs Wrong Recommendations*: To answer RQ_1 the first step is to classify the predictions by the three approaches as correct or wrong. We considered a prediction as correct if it represents an exact match (EM) with the target (*i.e.*, the expected output). This means that for the *code-to-comment* task the model generated a comment identical to the one written by human reviewers, while for the *code & comment-to-code* task the model implemented a code change required by the reviewer exactly as the human contributor did. For each pair of technique and automated task, such a process resulted in the identification of the buckets of correct and wrong predictions reported in Table II (columns "# correct (%)" and "# wrong (%)"). While the EM metric has been used in the evaluation of the three techniques, we acknowledge that it has strong limitations, since it provides quite a strict definition of correctness. For example, an automatically generated natural language comment requesting the same code changes of the target comment with different wording is considered wrong. While this undermines a purely quantitative assessment of performance, in our study we use EM only as a mean to identify *candidate* correct and wrong predictions. The predictions will undergo a manual analysis which, for example, considers correct generated messages being semantically equivalent to those posted by the human reviewers.

Qualitative Analysis. The goal of the manual analysis was to characterize the type of code change the experimented techniques were (or were not) able to request (*code-to-comment* task) or to automatically implement (*code & comment-to-code* task). For each of the above-described buckets of correct and wrong predictions (as identified via the EM analysis), we targeted the manual inspection of 167 valid instances, corresponding to a statistically significant sample with at least a confidence level of 99% and confidence interval of $\pm 10\%$ for each bucket. The target of 167 instances was defined by computing a sample size (SS) calculation formula [38] on the bucket having

the largest “population” (*i.e.*, wrong predictions generated by T5CR for the *code-to-comment* task, with 16'426 instances):

$$SS = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + \left(\frac{z^2 \times p(1-p)}{e^2 \cdot N} \right)}$$

where p is the predicted probability of the observation event to occur, set to 0.5 when not known a priori (as in our case), N is the population size, e is the estimated margin of error ($\pm 10\%$), z is the z -score for a given confidence level (in our case, 2.58 for the 99% confidence). As it can be seen from the formula, the larger N , the larger the sample size. Thus, using the largest “population” to compute the number of instances to inspect is a conservative choice, ensuring even better confidence when working on smaller buckets.

We use the term “valid” instances to account for the following scenarios. First, when inspecting a prediction falling in one of the *wrong* buckets it is possible that we realize that the prediction is actually correct (*e.g.*, the comment generated/retrieved by the technique uses different wording as compared to the target, but it is semantically equivalent). In this case, while we inspected a *wrong* instance, it will actually fall into the corresponding *correct* bucket. Second, we discarded several problematic instances we found in the test sets of the experimented techniques. For example, we found instances in the *code & comment-to-code* task for which, given the input code as context, it was impossible even for a human to understand the associated reviewer’s comment (*i.e.*, what the reviewer was asking to change in the input code). Indeed, the reviewer’s comment referred to a wider context (*e.g.*, parts of the code base not provided as input to the model), making the prediction impossible for the automated technique. These are problematic instances in the test set rather than failure cases of the technique, and we document them in RQ₂. In summary, an instance was considered “valid” if, given the input information (*i.e.*, C_s for the *code-to-comment* task, and $\langle C_s, R_{nl} \rangle$ for the *code & comment-to-code* task), it was possible for a human to understand the rationale for the related output: For the *code-to-comment* task, this means that the human evaluator was able to understand what the R_{nl} comment to generate refers to (*i.e.*, what the problem in the submitted code C_s spotted by the reviewer is); for the *code & comment-to-code* task, the evaluator considers an instance valid if the changes resulting in the revised code C_r to generate actually address the reviewer comment R_{nl} posted for the submitted code C_s . The instances to inspect were randomly selected from each bucket until the target number of valid instances was reached.

The columns “Inspected correct” and “Inspected wrong” in Table II report, for each bucket, the number of instances we ended up manually inspecting to reach our target of 167 valid instances per bucket. Overall, we inspected 2,291 instances. Each instance has been independently inspected by two authors (from now on, evaluators) who were tasked with classifying the type of change to request (*code-to-comment*) or to implement (*code & comment-to-code*). Five authors were involved in the manual analysis. On average, the authors have 13.4 years of programming experience (min = 6) and 9.2 years of experience with Java (min = 5), the language used in the inspected datasets.

One of them holds a PhD in software engineering, and three more are currently pursuing a PhD in software engineering. One is a software engineer.

The whole process was supported by a web app we developed that implemented the required logic and provided a handy interface to visualize the instance to label. For each instance, the evaluator was presented with: (i) the input provided to the approach; (ii) the generated prediction; and (iii) the expected output. As a result of the inspection, the evaluator could classify the instance or discard it as non-valid, providing an explanation as to why it was discarded.

The classification required to assign the instance one or more labels describing the change (*e.g.*, *refactoring* $\rightarrow$ *extract method*). Each evaluator was free to define their own label(s) (*i.e.*, open coding procedure), as they felt it was needed to properly describe the change: For this specific task, it was not possible to define upfront all possible labels, making card sorting [13] not suitable for our study. Indeed, while there are taxonomies of issues identified during the code review process [10], [24], [33] their abstraction level is not suitable for our goal. For example, the taxonomy by Mäntylä et al. [24] includes a category named *evolvability defects* $\rightarrow$ *structure* which is too coarse grained to investigate the automation capabilities of the subject techniques. To provide a concrete example, the taxonomy depicting the types of changes that the three techniques were (or were not) able to automatically request in the *code-to-comment* task (Fig. 1) we have an entire tree dedicated to the recommendation of refactoring operations (which would fall under the *evolvability defects* $\rightarrow$ *structure* taxonomy from [24]). Our taxonomy includes concrete refactoring operations (*e.g.*, *Extract method*, *Change variable/constant type*), some of which are successfully recommended by the experimented techniques, while others consistently represent failing scenarios. The fine-grained nature of our taxonomy allows to observe these differences.

The agreement among the authors was to define each label in the form *parent* $\rightarrow$ *child*, where *parent* was a coarse-grained description of the change while *child* was a more specific, fine-grained description. New labels defined by an evaluator were made available in the web application to the other evaluators. While this goes against the notion of open coding, this allows to reduce the chance of multiple evaluators defining similar labels to describe the same type of change while not substantially biasing the process. The evaluator was also in charge of flagging instances in the wrong buckets as “actually correct” in case they felt that the prediction, while different from the target, was correct.

The manual evaluation was performed in three rounds. A first round asked each evaluator to inspect 30 instances. This round resulted in a set of labels that has been inspected by the authors with the goal of merging similar labels and come up with a common strategy to categorize the instances in the next rounds. Then, a second round was performed in which the authors targeted the labeling of 30% of the overall instances assigned to them. Again, such a round was followed by a further inspection of the defined labels, with grouping of similar labels and further discussion on strategies to improve agreement. The rationale for the number of instances to inspect in each of the three rounds

was the following. We wanted to label very few instances in the first round (30) since we expected several inconsistencies in the way in which the authors were going to perform the labeling and, thus, we targeted a short dry run to test the adopted web application, the clarity of the overall process and, at least in part, the type of labels assigned by the authors (e.g., their granularity). Then, we decided to follow with a larger second batch (30%) which allowed to spot more corner cases worth to be discussed among the authors (e.g., instances for which an author was unsure about the type of label to assign). Finally, since we felt that the labeling process was well-defined and clarified among the authors, we decided to move on labeling the whole dataset. Once all 2,291 instances have been inspected by two evaluators, we solved conflicts that arose in 1,225 cases². Such a number may look high, since it represents 53% of the inspected instances. However, the high rate of conflicts is explained by three design decisions. First, we considered all conflicts, also the ones resulting in the first and second round in which the set of possible labels was not stable at all. Second, the labels in our study emerged from the data and were not pre-defined. To get an idea of the complexity of this task the whole process resulted in a total of 120 different labels. Third, we were conservative in our definition of conflict, which occurred if: (i) two evaluators assigned a different set of labels to the instance, even if the two sets partially overlapped; (ii) two evaluators assigned the exact same set of labels to a *wrong* instance with only one of the two reporting the instance as “actually correct”; (iii) only one of the two evaluators labeled the instance, while the other one discarded it. Each conflict has been inspected by two additional authors, who discussed and solved it.

Finally, we used the assigned labels to build hierarchical taxonomies showing the types of changes in which the three techniques tend to succeed and fail for the two automated tasks. Such a process required additional inspections of the considered instances. Indeed, once all categories in the taxonomies have been defined, two of the authors rechecked that all instances were assigned to the most proper category. Indeed, it is possible that a category C added during the very last labeling round would be more suitable for instances inspected at the very beginning of the manual process, when C was not available (since no one came up with that label while inspecting the instance). This resulted in the re-assignment of 16 instances ($\sim 1\%$).

The final number of valid instances (i.e., non-discarded) within each bucket is reported in the columns “Valid correct” and “Valid wrong” in Table II. A few clarifications are needed for values which are different from 167, which was our original target. First, we did not have correct (EM) predictions generated by CODEREVIEWER [23] for the *code-to-comment* task. Thus, we applied the following procedure to collect instances for the corresponding bucket. We selected among the *wrong* predictions generated by CODEREVIEWER, the top-100 in terms of BLEU-4 (Bilingual Evaluation Understudy) score [32]. BLEU measures how similar the candidate (predicted) and

reference (oracle) comments are in terms of overlapping 4-grams. A value of 1.0 indicates that the candidate and the predicted comment are identical. The selected top-100 predictions have a BLEU-4 ranging between 0.28 and 0.72.

Our assumption is that *wrong* predictions having a high BLEU score are likely to be correct, since they closely resemble the target comment. During the manual analysis, we discarded the instances that despite the high BLEU score, were actually wrong, since they did not belong to the “correct bucket”. This process led us to 50 valid instances in this bucket, which is the only one being underrepresented (see Table II). Second, as it can be seen from Table II, in several buckets we collected more than the targeted 167 valid instances. This is due to the conflict resolution phase in which some instances discarded by one of the two evaluators were considered valid and re-introduced.

The output of this analysis are two taxonomies reporting, for each of the two tasks, the types of code changes on which the experimented techniques tend to succeed or to fail.

Quantitative Analysis. We contrast the complexity of the test set instances resulting in correct and wrong predictions of the experimented techniques. For the *code-to-comment* task, we used as proxy for complexity the number of words featured in the comment to generate, under the assumption that longer comments are likely more complex.

For the *code & comment-to-code* task, we measure the number of AST-level changes required to convert C_s (i.e., the code submitted for review) into C_r (i.e., the revised code addressing the reviewer’s comment). We expect a higher number of changes to indicate a higher complexity of the comment to implement. The AST-level changes have been extracted using Gumtree Spoon AST Diff [14].

For the *code-to-comment* task, we used as proxy for complexity (i) the number of words featured in the comment to generate, under the assumption that longer comments are likely more complex, and (ii) the number of AST-level changes required to address the reviewer’s comment (as done for the *code & comment-to-code* task). The latter was only computed for the predictions generated by T5CR and by COMMENTFINDER, since for CODEREVIEWER we did not manage to retrieve from the dataset the code implementing the required change, but only the submitted code with the posted reviewer’s comment.

For both tasks, we report boxplots of the distribution of the complexity proxies for correct and wrong predictions both overall and by approach. We also statistically compare the two distributions assuming a significance level of 95% and using the Wilcoxon test [51]. The Cliff’s Delta (d) is used as effect size [16], and it is considered: negligible for $|d| < 0.10$, small for $0.10 \leq |d| < 0.33$, medium for $0.33 \leq |d| < 0.474$, and large for $|d| \geq 0.474$ [16]. We adjust p -values using Holm’s correction procedure [17]. We compare the complexity proxies only on the predictions we manually validated. The reason is that, as explained, relying on EM to identify correct predictions could lead to false negatives, thus invalidating our quantitative analysis.

2) *RQ2: Datasets Quality:* The datasets used to train and test the experimented techniques have been automatically mined from GitHub. The authors applied a number of heuristics to

²Given the open nature of the coding, it was not possible to compute a meaningful inter-rater agreement (e.g., Cohen’s kappa).

filter-out problematic instances. For example, in the *code-to-comment* task efforts have been made to remove review comments posted by bots. Similarly, in the *code & comment-to-code* task in which $\langle C_s, R_{nl} \rangle \rightarrow C_r$ triplets are involved, checks are performed to make sure that C_r (the code which should implement the reviewer's comment R_{nl}) is different from C_s . Indeed, $C_s = C_r \implies R_{nl}$ not addressed.

Despite the effort in cleaning the datasets, in our manual analysis we found $\sim 25\%$ of the inspected instances representing noise in the datasets, posing questions on the reliability of the evaluations reported in the literature. As previously explained, when discarding an instance during the manual analysis the evaluators had to report the reason why said instance was discarded. After such a process, the five authors looked at the provided motivations and distilled them into four main categories representing errors introduced during the automated mining of the data from GitHub. We answer RQ₂ by presenting statistics summarizing such an analysis.

3) RQ₃: *Comparison With LLMs*: We assess the performance of ChatGPT [1] on the two tasks focus of our study. We limited the number of samples to 250 for ChatGPT, due to the high cost of running such an evaluation. Indeed, there were two manual steps to perform. First, we needed to interact with the ChatGPT GUI to manually prompt each instance on which we wanted to run ChatGPT. Based on some tests we performed, we ended up selecting the following two prompts for our tasks:

code-to-comment: Write a code review of the following code "{inputCode}".

code & comment-to-code: Revise this code "{inputCode}" given this comment "{inputComment}".

In the prompts {inputCode} and {inputComment} represent the C_s and R_{nl} , respectively, in the test datasets used for the two tasks. The replies provided by ChatGPT when using these prompts made it clear that it properly interpreted the task to perform.

Second, once ChatGPT generates an answer, we cannot rely on EM to check if it is correct, since ChatGPT has not been trained to generate answers in the same format used in the test sets. For this reason we had to manually inspect the generated answers to assess their correctness. Each generated answer was independently inspected by two authors, who classified it as correct or wrong. For the *code-to-comment* task we considered the code review generated by ChatGPT as correct if it contains the target comment. For *code & comment-to-code*, we verified whether ChatGPT properly addressed the reviewer's comment, even with the coding solution being different to the target one. Conflicts (*i.e.*, the two authors disagreed on the correctness of ChatGPT's answer), which arose in 18% of cases, were solved by a third author.

For the *code-to-comment* task we randomly selected 50 instances from the test set of each of the experimented techniques (150 instances in total). The 50 instances included 25 correct (*i.e.*, the corresponding technique generated a correct solution) and 25 wrong predictions. The same approach has been used for the *code & comment-to-code* task which, however, is only automated by two of the three subject techniques, thus resulting in 100 randomly selected instances.

We answer RQ₃ by reporting the percentage of cases in which ChatGPT was successful in both tasks. We also analyze the overlap between the state-of-the-art techniques and ChatGPT by reporting the percentage of cases in which (i) both succeed; (ii) at least one of the two succeeds; and (iii) none of the two succeeds.

IV. RESULTS DISCUSSION

We discuss the achieved results by RQ. We use the  icon to mark lessons learned and directions for future work.

A. RQ₁: *Correct vs Wrong Recommendations*

Fig. 1 and Fig. 2 report the taxonomies of "types of code changes" involved in the predictions generated by the subject techniques. Fig. 1 refers to the *code-to-comment* task, depicting types of changes that the techniques were supposed to ask for in generated comments, as a human reviewer would do. Fig. 2 refers to the *code & comment-to-code* task, reporting types of changes that the techniques were required to automatically implement.

The taxonomies include several trees, each one representing a generic set of code changes specified in the root category (*e.g.*, *refactoring*, *bug-fix*). The number on the top-right corner of each label reports the number of instances we manually assigned to that change type (*e.g.*, 503 refer to *refactoring* in Fig. 1).

Three clarifications are needed on this point. First, one instance we analyzed (*i.e.*, a prediction generated by an approach for a test entry) could have been assigned to multiple categories since requiring multiple types of changes. Second, for readability reasons, we decided to not report in the picture all categories of code changes that have been assigned to less than ten instances. Indeed, it is difficult to draw any conclusion with so few data points. The full data is available in our replication package [4].

Third, being a hierarchical taxonomy, one may expect the number of elements in a parent category to match the sum of the number of elements in its child categories. However, this is not the case due to two reasons. First, sometimes the parent category has been used as a label when we did not manage to clearly identify the required code change, but only its overall goal (*e.g.*, using *bug fix* $\rightarrow$ *fix wrong behavior* instead of its child *modify if condition*). Second, as previously explained, we do not depict in the taxonomies categories featuring less than 10 instances. However, we still count their instances in the parent category (*e.g.*, *refactoring* $\rightarrow$ *renaming* has a *rename class* child which has not been depicted but contributes with 6 instances).

In this scenario, the parent category has been used as a label when we did not manage to clearly identify the required code change, but only its overall goal (*i.e.*, fixing wrong behavior). This is another reason why parent categories can have a higher counting than the sum of their child categories.

The color assigned to each label reflects the ability of the techniques to automate the code review task in the context of such a change type. Since we manually analyzed $\sim 50\%$ of correct and $\sim 50\%$ of wrong predictions generated by each

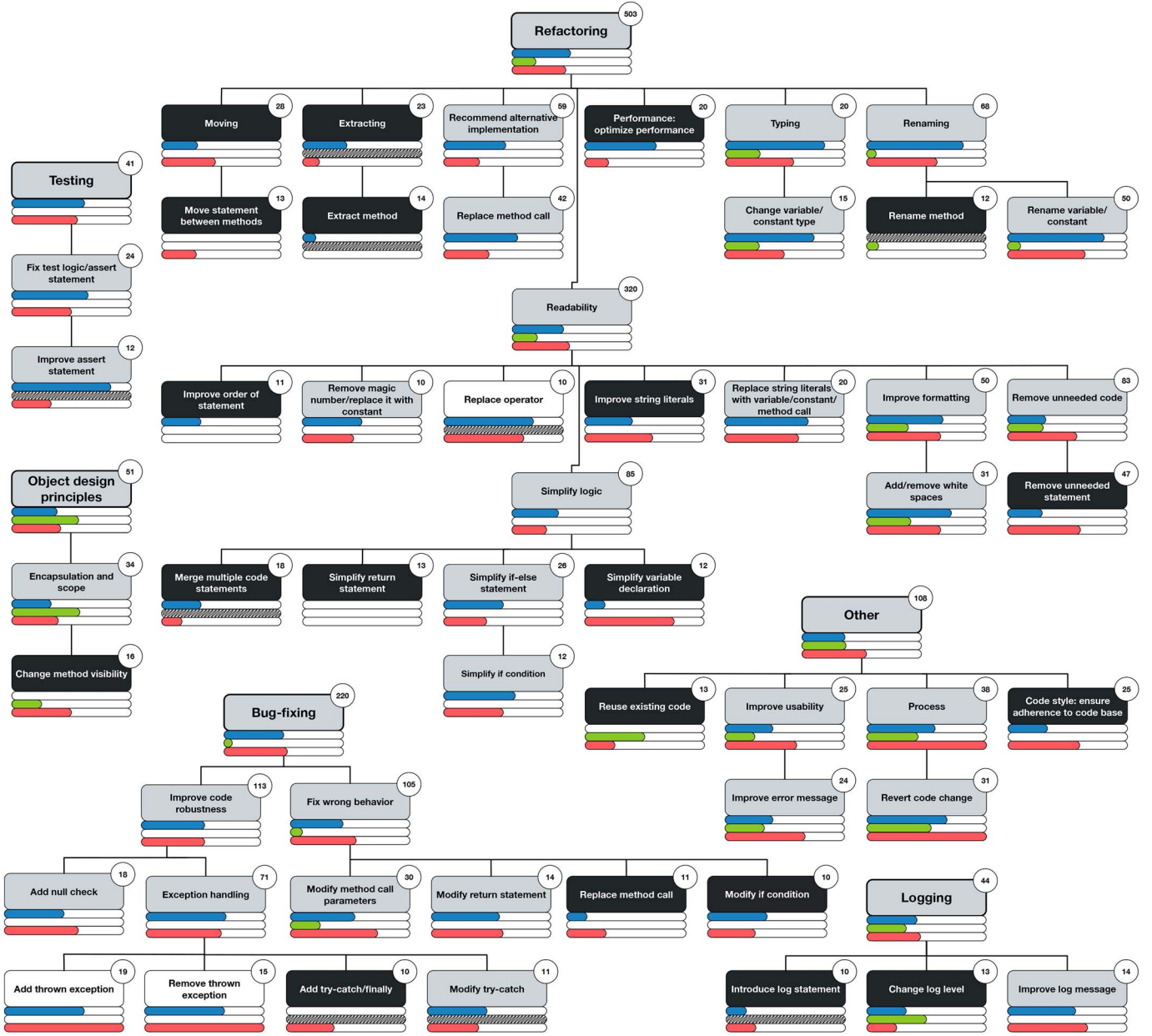


Fig. 1. Taxonomy of types of changes for the *code-to-comment* task. The color assigned to each label reflects the ability of the techniques to automate the code review task in the context of such a change type (white best, black worst). We report the percentage of successful predictions by each approach for each change type as bars below the corresponding category: T5CR (blue bar), CODEREVIEWER (green), and COMMENTFINDER (red).

approach, a success rate around 50% for a change type indicates that the techniques do not tend to perform particularly well or bad for that change type.

Indeed, the goal of this analysis is to see if the correct (wrong) predictions are polarized by specific categories. For these reasons, we defined the color schema as follows:

A gray label indicates a change type for which the automation level aligns with what expected based on our sample of correct and wrong predictions. Looking at Table II it can be seen that, for the *code-to-comment* task, we inspected a total of 388 correct and 544 wrong instances. Such an imbalance is due, as previously explained, to the CODEREVIEWER technique

which generated 0 EMs and for which we decided to manually inspect 100 wrong predictions looking for actually correct ones (50 identified). This means that we should expect an average performance per change type close to $(388 * 100) / (544 + 388) = 42\%$. For this reason, Fig. 1 features a grey category if the techniques succeeded for such a change type in 32% to 52% of cases (*i.e.*, $42\% \pm 10\%$). With a similar computation, Fig. 2 features a grey category if the techniques succeed in 43% to 63% of cases (since an average of 53% correct predictions has been analyzed per approach). Note that the “ $\pm 10\%$ choice” has been done to simplify the results discussion and visualization. We acknowledge that other choices are possible (*e.g.*, $\pm 20\%$);

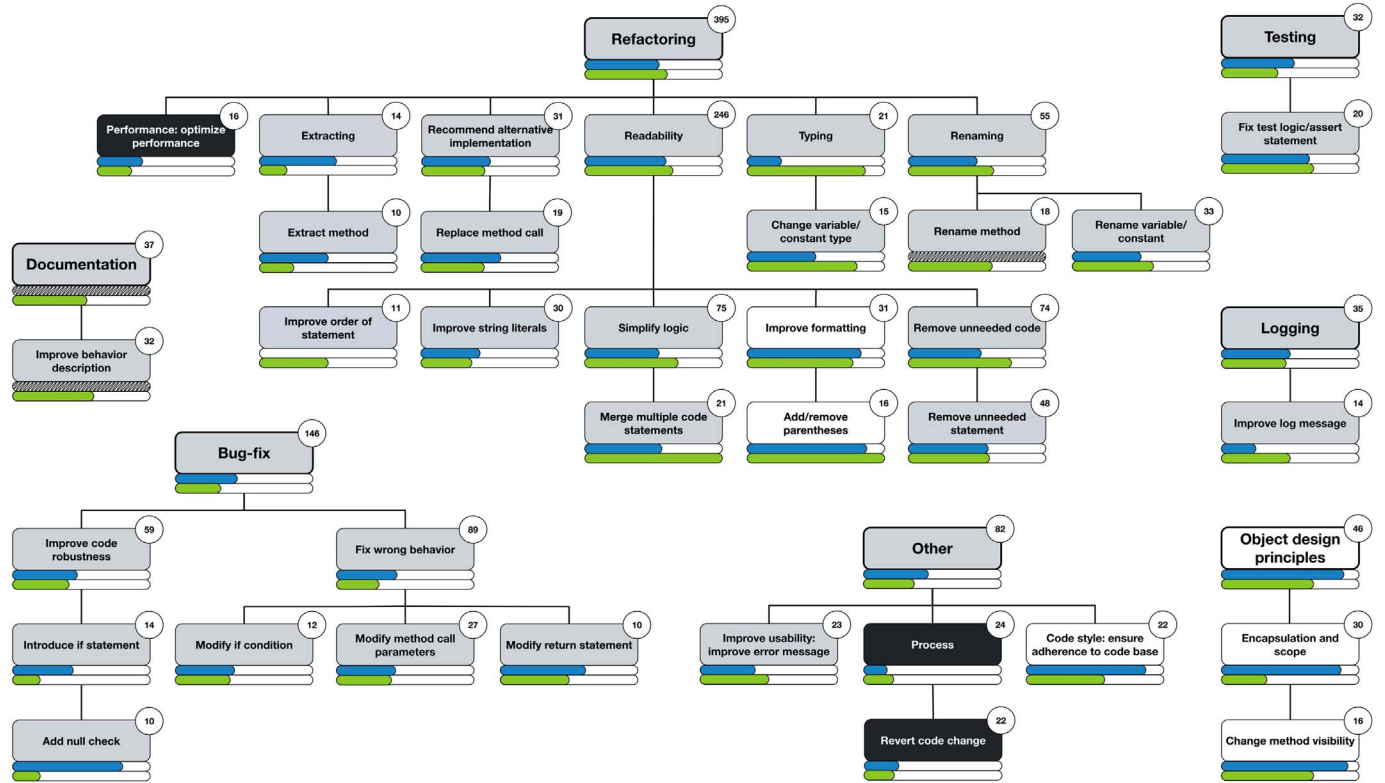


Fig. 2. Taxonomy of types of changes for the *code & comment-to-code* task. The color assigned to each label reflects the ability of the techniques to automate the code review task in the context of such a change type (white best, black worst). We report the percentage of successful predictions by each approach for each change type as bars below the corresponding category: T5CR (blue bar), CODEREVIEWER (green).

raw data with exact percentages are available in our replication package [4].

A **black** label indicates a change type for which the automation tends to fail, thus in which the techniques are struggling. This means a success rate lower than 32% for the *code-to-comment* task, and lower than 43% for the *code & comment-to-code* task.

A **white** label indicates a change type for which the automation succeeds more than expected, namely in at least 53% of cases for *code-to-comment* and 64% of cases for *code & comment-to-code*.

While this 3-level color schema represents the capabilities of the experimented techniques as a whole, Fig. 1 and Fig. 2 also report the percentage of successful predictions generated by each approach for each change type as bars below the corresponding category. In Fig. 1 the three bars are ordered from top to bottom as: T5CR (blue bar), CODEREVIEWER (green), and COMMENTFINDER (red). In Fig. 2 the bars are only two, corresponding to T5CR (blue) and CODEREVIEWER (green). An empty bar indicates that the approach always failed for instances of that type. Instead, the filling of the bar with a zig-zag pattern indicates that the manually inspected test set entries on which the corresponding technique has been experimented did not contain any instance of that type.

On top of the two taxonomies, Fig. 3 depicts the boxplots showing the computed “complexity” of the test instances on

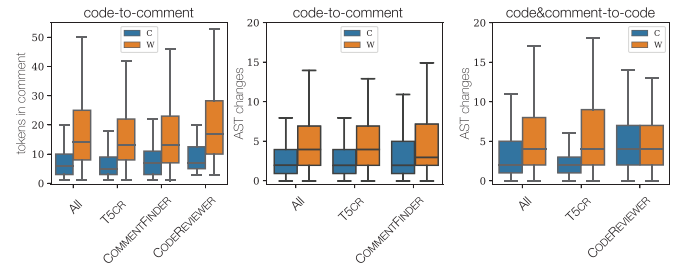


Fig. 3. Task complexity for correct and wrong predictions.

which the experimented techniques succeed (blue) or fail (orange). We discuss our qualitative and quantitative results by automated task.

1) *Code-to-Comment*: Before looking at characteristics of correct and wrong predictions, Table III reports, for each approach and for each task, the percentage of non-EM predictions that we classified as actually correct. As it can be seen, 21.83% of non-EM predictions generated by CODEREVIEWER are actually correct for the *code-to-comment* task. The percentage is smaller for the other two techniques for which we did not focus the inspection on predictions having a high BLEU, but still non-negligible ($\sim 2.5\%$). For example, in the case of COMMENTFINDER the EM predictions are 2.85% of the test set instances, while we found an additional 2.22% of non-EM predictions which are actually correct, almost doubling

TABLE III
NON-EM CLASSIFIED AS CORRECT DURING MANUAL ANALYSIS

Reference	Task	% Correct Non-EM
T5CR	<i>code & comment-to-code</i>	15.66%
	<i>code-to-comment</i>	2.58%
COMMENTFINDER	<i>code-to-comment</i>	2.22%
CODEREVIEWER	<i>code & comment-to-code</i>	17.37%
	<i>code-to-comment</i>	21.83%

the approach’s correctness. A manual analysis of (a sample of) non-EM predictions is needed to better assess the capabilities of an automated technique. An example of non-EM generated by CODEREVIEWER and being actually correct belongs to the *other* → *reuse existing code* category: The target comment was “Use `IOUtils` instead”, while CODEREVIEWER generated the equivalent “Can we use Guava’s `IOUtils` here?” This is a first important outcome of our study: ♡ Using EM to assess the automation of the *code-to-comment* task might be unfair.

Not surprisingly the white categories (*i.e.*, the techniques tend to succeed) are characterized by simple requests to include in the generated message, and in particular the *removal/addition of a thrown exception*, and the *replacement of an operator*. The excellent performance achieved in these change categories are usually driven by the success of the COMMENTFINDER IR-based technique. The latter has 100% accuracy in recommending the addition/removal of exceptions, thanks to the retrieval from the training set of past reviewers’ comments requiring such a change for similar methods. ♡ Looking at the taxonomy, it is clear that for code change types which are quite general, simple, and thus likely to be requested over and over again in different code review instances (*e.g.*, the addition/removal of exceptions, asking to *revert a code change*), an IR-based approach can be a trump card. Differently, comments requiring the description of more complex changes are challenging to retrieve or synthesize. The *refactoring* tree provides interesting examples.

Simple refactorings such as *changing variable/constant type* or *renaming variable/constant* are overall well-supported (*e.g.*, COMMENTFINDER: “`qry` → `query`”). When it comes to refactorings involving complex code changes, possibly impacting multiple code components, the techniques tend to fail. This is the case for refactorings *extracting* or *moving* code elements. This is likely due to the limited contextual information provided to code review automation techniques. Among the experimented techniques, T5CR and COMMENTFINDER work at method-level granularity, meaning that the method provided as input represents everything the model knows about the system. Similarly, the “view” of CODEREVIEWER is limited to a diff hunk. Such an issue does also affect the performance of the techniques for apparently simple changes to require, such as those asking to *change the method visibility*. Indeed, without additional knowledge of the system it is difficult to judge what the correct visibility of a method should be. ♡ Pushing the boundaries of code review automation for these types of

changes requires enriching the contextual information provided to the techniques, similarly to what observed for other software engineering tasks [44].

A negative exception in the refactoring tree is the *renaming of methods* which one could expect to be on a similar level of difficulty as compared to the *renaming of variable/constants* which is, instead, quite successful. We inspected these instances and we noticed that, while renaming variables/constants may just require a term expansion (as in the `qry` example previously reported), methods’ names are more expressive and challenging and, while the techniques are sometimes able to capture the need for a renaming, they fail in recommending meaningful alternatives.

The techniques also have a hard time automating logging activities, especially when it comes to suggest the *introduction of a log statement* or the need to *change the log level* (*e.g.*, from error to warning). Interestingly, for both method renaming and recommendation of log-related changes, specialized DL-based techniques have been proposed in the literature (see *e.g.*, Alon et al. [6] for renaming, and Mastropaolo et al. [25] for logging). Based on the reported empirical evaluations, those techniques proven to be quite effective in these tasks. For example, Mastropaolo et al. [25] presented LANCE, an approach able to correctly recommend fixes to the level of log statements in 66% of cases. ♡ While the code review automation techniques proposed in the literature target the automation of generic code changes, the adoption of specialized techniques might be more effective for specific change types. However, this might not be straightforward to do in the context of the *code-to-comment* task. Indeed, assuming the will to specialize a model for “commenting” on a specific type of issue, the first needed ingredient is a training dataset, which might not be easy to collect. One may think to cluster reviewers’ comments via lexical analysis and train a specialized model on each of those clusters. Nevertheless, a trade-off between cohesiveness of the clusters and availability of training data soon becomes evident: very cohesive clusters will result in highly specialized models which, however, are likely to benefit from very little training data (*e.g.*, only a few instances in which a reviewer’s comment is suggesting to introduce a log statement). Larger clusters featuring more training data are instead unlikely to specialize the model for specific types of recommendation, thus again pushing it towards a generic recommender. A more promising approach might be to manually define “commenting patterns” for a specific type of change (*i.e.*, a standard sentence expressing the need for improving a certain aspect of the code, such as introducing a log statement). In this case, the training dataset could be built by parsing code changes performed during the change history of a project (*e.g.*, commits introducing a new log statement), independently from the availability of code review information for these changes. This implies the possibility to reliably identifying code changes in which the target issue has been fixed. For some of the “black categories” in our taxonomy this can be easily achieved (*e.g.*, lack of log statement, need for changing the method visibility, etc.). Others would require more advanced solutions, like the usage of tools to detect refactoring operations [45]. Negative instances, *i.e.*, code components on

which the target issue does not manifest (*e.g.*, no need to add log statements), might be needed as well. Once trained, specialized models can be triggered on the code change submitted for review, reporting the improvement recommendations (if any) to the developer.

Not surprisingly, the experimented techniques do not shine in recommending types of code changes which are likely to be system-specific and, thus, difficult to learn/retrieve from other sources. This is the case for the *performance optimizations* comments that the techniques were required to emulate (*e.g.*, TARGET: “We should use `keyService` here, intention is to cache key temporary so under heavy load we don’t download keys all the time”). ♡ A possible strategy to overcome this limitation could be to fine-tune the techniques to a specific software project or, at least, to a set of projects falling in the same domain (*e.g.*, DB engines). For example, after a pre-training performed on generic code review data, a DL-based approach could be fine-tuned to specifically support code review in a project. A major obstacle is the possible lack of fine-tuning training data, since a single project is unlikely to provide enough training instances. This may be partially overcome through data augmentation techniques [55].

T5CR and COMMENTFINDER achieve good performance when it comes to asking for *testing*-related changes, correctly generating comments aimed at both improving the test coverage/logic (*e.g.*, T5CR: “add a check here to verify that the `serialDataReceived` method was not called”) and cleaning/refining them, *e.g.*, T5CR suggested to replace an empty `String` passed as parameter in an assert statement with an `EMPTY_VALUE` constant already used in other statements of the test. In general, the changes to recommend in the *testing* category are very specific and tend to focus on a single code statement. ♡ Still, this shows the potential of these techniques as possible “refinement tools” for approaches supporting the automated generation of test cases [15], [31].

From the quantitative perspective, the boxplots in Fig. 3 for the *code-to-comment* task suggest, as expected, that the techniques tend to succeed in the generation of simple reviewers’ comments, having a median of 6 words composing them. As a comparison, the failing cases are more than twice longer (in terms of median), with 14 words. Such a trend holds, with minor differences, for all approaches. Also, it is interesting to notice that, when considering all techniques together, the first quartile of the wrong predictions is very close to the third quartile of the correct predictions, indicating a strong difference between the two sets that is confirmed by the statistical analysis with a p -value < 0.001 accompanied by a large effect size (test results in [4]). Similar observations can be drawn when using the AST changes required to implement the reviewer’s comment as a complexity proxy. ♡ Future work should focus on boosting performance in these challenging scenarios, since the approaches seems to work pretty well in the generation of simple comments (<20% of wrong predictions have less than 6 words).

2) *Code & Comment-to-Code*: Also for this task, we start by observing the substantial percentage of non-EM predictions which are actually correct — 15.66% for T5CR and 17.37%

for CODEREVIEWER. ♡ This reinforces the need for manual analysis when assessing the performance of techniques for code review automation.

The taxonomy depicted in Fig. 2 is smaller as compared to the previous one, since a higher number of categories (91) count less than 10 instances (for the full taxonomy see [4]). It is interesting to see some major differences as compared to the previous taxonomy. Changes which were trivial to ask for in a comment to generate (*e.g.*, “please revert this change”) are challenging to automatically implement, as required in the *code & comment-to-code* task. Indeed, the reverting may require several code changes which are not necessarily easy to predict, especially if the full code diff is not part of the information available to the model.

Interesting is the complementarity between the two techniques that support this task. T5CR is very effective in changes related to *object design principles*, *e.g.*, handling issues related to *encapsulation and scope* of variables/methods, which usually require minor code changes. Also, T5CR works well in implementing changes *ensuring adherence to the code base* in terms of *coding style*, *e.g.*, addressing comments like “`String.isEmpty()` is used in other places” pointing to an inappropriate *if* condition checking `coverageId.length() == 0`. CODEREVIEWER, instead, is the only approach supporting documentation changes, and can achieve excellent performance even for less-trivial code changes requiring *e.g.*, to *merge multiple code statements* (in order to improve readability) or to migrate towards more appropriate types (*refactoring* → *typing*). Considering that both techniques are built on top of a transformer-based architecture, such a complementarity can be partially explained by the different code representation they use, one looking at a single method at a time (T5CR) and one taking a diff hunk into account (CODEREVIEWER), possibly with a partial view of specific code components (*e.g.*, only the changed lines of a method are visible in the diff hunk). ♡ A hybrid representation including both the full representation of the involved code entities and the changed lines might help in getting the best of both worlds.

As expected, both approaches are effective in the automation of very simple changes related to *improve the formatting* of code (*e.g.*, *add/remove parentheses*, *add/remove white spaces*). The automation of these changes can be easily performed by a code formatter (*e.g.*, [3]), without the need for expensive DL models. ♡ Such instances should be removed from the test sets used in the evaluation of techniques automating the *code & comment-to-code* task, to avoid inflating the percentage of EM predictions they generate.

The two techniques struggle to automatically implement complex *bug-fixes* requiring major code changes (*e.g.*, *fix wrong behavior*) and/or changes to the code logic (*e.g.*, *modify if condition*). ♡ This result is inline with what observed for techniques specialized in automated bug-fixing [47], which also tend to be successful in a minority of cases (usually <30%) confirming the need for more work in the area.

Finally, we want to comment on the performance of the two techniques on the 87 types of code changes which are not represented in Fig. 2, since counting less than 10 instances each.

TABLE IV
CATEGORIES OF DISCARDED INSTANCES

Reason	#	T5CR		COMMENTFINDER	CODEREVIEWER	
		C&NL2C	C2NL	C2NL	C&NL2C	C2NL
Unclear comment	281	32	55	97	57	40
No change asked	182	24	13	44	30	71
Ignored comment	74	25	0	0	49	0
Wrong linking	16	1	0	1	4	10
Other	16	2	1	1	1	11

Overall these 87 categories feature 132 of the predictions we inspected for the *code & comment-to-code* task.

Out of those, 69 (52%) were correct predictions, which matches the expected success rate and seems to suggest that the two state-of-the-art techniques do not really struggle in automatically implementing code changes which are likely to be less represented in the training set. However, by inspecting the predictions in these categories, we found out that a “good” level of performance (*i.e.*, $\geq 52\%$ correct predictions) is only obtained for 48% of these poorly represented categories (*e.g.*, for 18 of them we observed 0% of correct predictions). It is thus possible that, in some cases, the models learn from similar and related categories in a sort of transfer learning fashion. For example, training on instances of the well-represented category “*remove unneeded statement*” might have played a role in achieving good performance on the five instances belonging to the code change type “*remove final modifier*”. However, given the low number of instances in each of these categories (at most 9), we cannot draw any conclusion based on these findings.

The results of the quantitative analysis (right side of Fig. 3) show an interesting trend: while the correct predictions by T5CR require substantially simpler changes as compared to the wrong predictions (median of 2 AST changes *vs* 4, p -value < 0.001 with medium effect size), this is not the case for CODEREVIEWER. Here the two sets are basically equivalent in terms of code change complexity (negligible effect size). Looking at the boxplots it is clear that T5CR tends to overfit to simpler changes, while CODEREVIEWER, possibly thank to the diff hunk representation, is able to cope with more complex code changes as well, supporting its status as state-of-the-art approach.

B. RQ₂: Datasets Quality

Table IV reports the number of instances that we discarded as being problematic together with the reason why they have been discarded. For the sake of space, we shortened the *code & comment-to-code* task as C & NL2C and the *code-to-comment* as C2NL. While Table IV shows also the results by test set of each technique, we focus the discussion on the overall trend (#). The “Other” category contains 16 instances discarded for various but rare reasons, that we do not discuss but document in [4]. The remaining “discarding reasons” are sorted based on their frequency from top to bottom.

For 281 cases, we assigned the *unclear comment* label to discard the instance since it was impossible even for a human to understand what to implement (*code & comment-to-code*)

or what the target comment to generate was actually asking to the developer (*code-to-comment*). For the *code & comment-to-code* task, this was due to the limited contextual information provided to the model (*i.e.*, the input). For the *code-to-comment* task, a recurring problem is again the lack of context but, this time, related to the conversation that happened between the contributor and the reviewer(s), which is not visible to the techniques. For example, a comment saying “ah, ok, that would be clearer” is meaningless without knowing the previous exchanged messages. ♡ As already observed in RQ₁, increasing the contextual information is a must to push the boundaries of code review automation.

In 182 instances we inspected the reviewer’s comment was not requesting any change. For the *code & comment-to-code* task, this means that the approaches could not really address the comment by modifying the code (*e.g.*, “Awesome work so far, Eli!”). For the *code-to-comment* task this means training and testing the technique for the generation of comments which are uninteresting given the automation goal. Indeed, these techniques aim at generating comments asking for code changes targeting the improvement of source code. Thus, comments like “I am not sure what GitHub wants to tell me with this icon here :)” should not be considered relevant for these approaches. The cleaning pipelines employed in the works presenting the experimented techniques fail in filtering out those meaningless instances.

In 74 cases (all related to the *code & comment-to-code* task), while the reviewer’s comment was asking for a change, the contributor was changing the code but not to address the comment. These instances penalize both the learning and the evaluation of the models. Indeed, even assuming that the model correctly implements the change required by the reviewer, during training the weights of the network will be revised to steer the prediction towards a different (wrong) target and, during evaluation, any quantitative metric is likely to point to a wrong prediction.

Finally, 16 instances result from errors while mining the dataset, since the code has been linked to a wrong code, *e.g.*, “there’s no need for `final` in interfaces” for an input code not being an interface.

♡ Worth noticing is the overall number of discarded instances (574) out of the 2,291 manually analyzed (25%). Since the test set is just a random selection of 10% of data, we can assume a similar distribution in the training sets of the subject techniques. Thus, all discussed instances have the potential to hinder the training and bias the testing, since it is unreasonable to expect them to result in a correct prediction given the target. Supervised or unsupervised techniques aimed at removing these problematic instances are needed to make a further step ahead on code review automation thanks to higher-quality datasets.

C. RQ₃: State-of-the-Art vs ChatGPT

Table V shows the results of the manual analysis assessing ChatGPT in automating code review: For each task (rows) we report the percentage of cases in which ChatGPT succeeds (✓) or fail (✗) for instances on which the state-of-the-art (SOTA) techniques were correct or wrong. Results are aggregated for

TABLE V
MANUAL ANALYSIS OF CHATGPT PREDICTIONS

Task	✓State-of-the-art		✗State-of-the-art	
	✓ChatGPT	✗ChatGPT	✓ChatGPT	✗ChatGPT
code & comment-to-code	66%	34%	44%	56%
code-to-comment	19%	81%	7%	93%

the three approaches, with raw data available in our replication package [4].

ChatGPT performs slightly better than the SOTA in the *code & comment-to-code* task, being able to address the reviewer's comment in $55/100 = 55\%$ of cases, as compared to the 50% of the three techniques (we selected half instances on which they work, and half on which they fail — see Section III.B.3). Interesting is the complementarity between ChatGPT and the SOTA: ChatGPT succeeds in 44% of cases in which the SOTA techniques fail. These are mostly instances in which the reviewer's comment provides little information about the change to implement. For example, ChatGPT addressed a reviewer's comment pointing to a "wrong formula" in the code ("wrong formula" is the full content of the reviewer's comment) by applying the following change: `forward * strikes` Like.get(i) + shiftOutput; → `forward * Math.exp(strikesLike.get(i)) + shiftOutput`; . This is a failing instance for the SOTA.

Concerning the *code-to-comment* task, ChatGPT succeeds in $19/150 = 13\%$ of cases, being less performant than the SOTA. Only 5 of those instances are failure cases for the SOTA. For this task the output of ChatGPT is not a single comment (as for the SOTA techniques) but a list of observations regarding the submitted code. In most of cases, we found these comments to be meaningful, but they often miss or are in contrast with the point raised by the human reviewer.

This is the case for an instance in which the diff hunk reported a change from `trace` to `info` for the level of a logging statement. While the reviewer complained about this change (and the SOTA technique agreed with the human reviewer), ChatGPT was in favor of it, commenting: "*an info level logging statement would be more applicable*".

Since there is evidence in the literature about the major role played by the provided prompt on the output produced by large language models when dealing with code-related tasks [26], we further investigated whether it is possible to improve the performance of ChatGPT by exploiting different and more advanced prompts. We took the failure cases of ChatGPT (*i.e.*, 131 cases for the *code-to-comment* task and 45 cases for the *code & comment-to-code* task) and we provided them again as input to ChatGPT, but with different promptings also exploiting information from our taxonomy. In particular, our taxonomies feature five root categories classifying the types of changes reviewers usually ask to implement in a code review process: *Bug-Fixing*, *Object-Design Principles*, *Testing*, *Refactoring*, *Logging*. On top of that there is the *Other* category grouping change types which cannot be attributed to any of the five above categories. We use these root categories to define two prompts (one for each of the two tasks), inspired by the chain-of-thought

prompting methodology [50] which has been successfully applied in natural language processing:

***code & comment-to-code* task**

Given this code "{inputCode}" and this comment "{inputComment}" and assuming you are an expert developer:

- 1) Identify/Classify the type of code changes requested in the comment. Answer with one or more of the following:
 - Changes have been required to refactor the code to improve its quality;
 - Changes have been required since tests for this code must be written;
 - Changes have been required to better align this code to good object-oriented design principles;
 - Changes have been required to fix one or more bugs;
 - Changes have been required to improve the logging of its execution;
 - Changes have been required for other reasons not listed above.
- 2) Implement the required code changes.textcbrc:

***code-to-comment* task**

Given this code "{inputCode}" and assuming you are an expert code reviewer:

- 1) Decide whether the code needs to be revised or not. Answer True or False.
- 2) If the response to the above point is True, then identify/classify the type of code change(s) required. Answer with one or more of the following:
 - Changes are needed to refactor the code to improve its quality;
 - Changes are needed since tests for this code must be written;
 - Changes are needed to better align this code to good object-oriented design principles;
 - Changes are needed to fix one or more bugs;
 - Changes are needed to improve the logging of its execution;
 - Changes are needed for other reasons not listed above.
- 3) Write a code review (*i.e.*, explain the changes to be performed, if any) based on your answers to the questions above.

As it can be seen, both prompts guide the large language model towards the solution of the task without, however, giving it any extra information besides what would be available during the review process.

As for the *code & comment-to-code* task, we also experimented an additional prompt, which classifies the reviewer's

TABLE VI
FAILURE CASES MADE SUCCESSFUL VIA MORE ADVANCED PROMPTING

Task	Prompt	✓ChatGPT (%)
<i>code & comment-to-code</i>	Chain-of-Thought	23/45 (51%)
<i>code & comment-to-code</i>	Label-Aware	25/45 (55%)
<i>code-to-comment</i>	Chain-of-Thought	13/131 (10%)

comment into one of the categories in our taxonomies before asking the model to automatically implement the code changes required to address the comment

Assume you are an expert developer. Given this code "{inputCode}" implement the code changes requested by a code reviewer in this comment "{inputComment}", considering that the change requested is asking to [category from our taxonomy to which the comment has been manually assigned].

For example, assuming that the assigned category in our taxonomy *logging*, the prompt will end with: considering that the change requested is asking to improve the logging of its execution.

In other words, we are simulating the scenario in which, besides providing a comment explaining the change to perform, the reviewer also provides a label classifying the type of code change required. Note that such a prompting cannot be experimented in the *code-to-comment* task, since in that case the reviewer comment is the output of the model rather than the input.

As previously done, two authors independently assessed the correctness of ChatGPT output, and conflicts have been solved by a third author. Table VI reports the achieved results. For the *code & comment-to-code* task, depending on the used prompt, we managed to make ChatGPT correctly predicting up to 25 of the previously identified 45 failure cases. It is quite interesting to see that just providing the model with an explicit label taken from our taxonomy and describing the type of change required in the reviewer's comment substantially helps the model in correctly implementing the required changes. As for the *code-to-comment* task, the chain-of-thought prompt helped the model only 13 of the 131 failure cases, confirming that generic large language models not specifically trained for this task struggles in dealing with it.

For completeness, we also provide in our replication package [4] two files mapping the ChatGPT success and failure cases to categories in our taxonomy. We do not discuss the findings in terms of coverage of the different categories in our taxonomy since the number of instances we analyzed in RQ₃ is quite low, resulting in few data points for each category (*i.e.*, it is difficult to judge whether CHatGPT works well for specific code change types).

💡 ChatGPT is a very competitive baseline for the *code & comment-to-code* task, with prompting playing a major role in boosting the model's performance. However, it is important to consider the fact that ChatGPT may have seen

the code addressing the reviewer's comment during training, questioning the extent to which such a comparison is fair. When it comes to generating reviewers' comments, SOTA techniques are superior, supporting the worth of further research in this direction.

V. THREATS TO VALIDITY

Construct validity. In RQ₁ we classified the change type based on what was visible in the test set instance. For example, we did not see the conversation between the contributor and the reviewers, which could allow to better understand the reviewers' requests. However, this was done by design to identify "unclear" instances in RQ₂.

Also, we considered the predictions of the three techniques on the test sets used in the original papers presenting them. This means that they have not been assessed on the same dataset. However, as previously explained, our goal was not to compare the capabilities of the three approaches, but rather to look at their strengths and weaknesses as representative of the state-of-the-art in code review automation.

Internal validity. There are possible subjectiveness issues in the manual analyses. We mitigated this threat by always involving multiple authors inspecting the same instance. Still, as for any manual process, imprecisions are possible.

External validity. In terms of studied techniques, we considered all those focusing on the automation of the two targeted tasks; we only excluded the approach by Li L. et al. [22] for the reasons explained in Section III.A.1. All our RQs required manual analysis, resulting in limitations on the number of data points collected. Given the expense of the manual investigation, we targeted the inspection of a sample of instances ensuring at least a confidence level of 99% and a confidence interval of $\pm 10\%$ on each of the analyzed bucket of predictions. More conservative choices (*e.g.*, $95\% \pm 5\%$ confidence) would have required a much higher number of inspected instances (more than twice). Replications are needed to corroborate/revise our findings.

VI. CONCLUSION AND FUTURE WORK

We assessed the capabilities of three state-of-the-art techniques for code review automation [18], [23], [48]. Differently from the mostly quantitative evaluations available in the literature our study has a strong qualitative focus. Our study disclosed the scenarios in which state-of-the-art approaches tend to succeed and fail (RQ₁) and identified major issues in the quality of the datasets used for their training and evaluation (RQ₂). Finally, we showed that ChatGPT [1], as representative of Large Language Models, is a competitive technique for code review automation, but still struggles in several scenarios, justifying the need for more research on models specialized for such automation (RQ₃).

Future work will focus on assessing (i) the usefulness of code review automation from the practitioners' perspective, and (ii) the code review automation opportunities offered by other state-of-the-practice techniques (*e.g.*, Copilot [2]).

REFERENCES

- [1] "ChatGPT," OpenAI. Accessed: Mar. 27, 2023. [Online]. Available: <https://openai.com/blog/chatgpt>
- [2] "Copilot website," GitHub. Accessed: Nov. 10, 2022. [Online]. Available: <https://copilot.github.com>
- [3] "Prettier," Prettier. Accessed: Mar. 25, 2023. [Online]. Available: <https://prettier.io>
- [4] GitHub. Accessed: Nov. 2023. [Online]. Available: https://github.com/CodeReviewAutomationSota/code_review_automation_sota
- [5] W. H. A. Al-Zubaidi, P. Thongtanunam, H. K. Dam, C. Tantithamthavorn, and A. Ghose, "Workload-aware reviewer recommendation using a multi-objective search-based approach," in *Proc. 16th ACM Int. Conf. Predictive Models Data Analytics Softw. Eng. (PROMISE)*, 2020, pp. 21–30.
- [6] U. Alon, M. Zilberstein, O. Levy, and E. Yahav, "Code2vec: Learning distributed representations of code," in *Proc. ACM Program. Lang.*, vol. 3, no. POPL, pp. 40:1–40:29, 2019.
- [7] S. Asthana et al., "WhoDo: Automating reviewer suggestions at scale," in *Proc. 27th ACM Joint Meeting Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng. (ESEC/FSE)*, 2019, pp. 937–945.
- [8] A. Bacchelli and C. Bird, "Expectations, outcomes, and challenges of modern code review," in *Proc. Int. Conf. Softw. Eng.*, Piscataway, NJ, USA: IEEE Press, 2013, pp. 712–721.
- [9] G. Bavota and B. Russo, "Four eyes are better than two: On the impact of code reviews on software quality," in *Proc. 31th IEEE Int. Conf. Softw. Maintenance Evolution (ICSME)*, 2015, pp. 81–90.
- [10] M. Beller, A. Bacchelli, A. Zaidman, and E. Juergens, "Modern code reviews in open-source projects: Which problems do they fix?," in *Proc. 11th IEEE/ACM Work. Conf. Mining Softw. Repositories (MSR)*, 2014, pp. 202–211.
- [11] A. Bosu and J. C. Carver, "Impact of peer code review on peer impression formation: A survey," in *Proc. 7th IEEE/ACM Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, 2013, pp. 133–142.
- [12] M. Chouchen, A. Ouni, M. W. Mkaouer, R. G. Kula, and K. Inoue, "WhoReview: A multi-objective search-based approach for code reviewers recommendation in modern code review," *Appl. Soft Comput.*, vol. 100, 2021, Art. no. 106908.
- [13] A. Coxon, *Sorting Data: Collection and Analysis* (Quantitative Applications in the Social Sciences), Newbury Park, CA, USA: SAGE Publications, 1999, no. 127.
- [14] J. Falleri, F. Morandat, X. Blanc, M. Martinez, and M. Monperrus, "Fine-grained and accurate source code differencing," in *Proc. 29th IEEE/ACM Int. Conf. Automated Softw. Eng. (ASE)*, 2014, pp. 313–324.
- [15] G. Fraser and A. Arcuri, "EvoSuite: Automatic test suite generation for object-oriented software," in *Proc. 21st ACM Joint Meeting Eur. Softw. Eng. Conf. ACM/SIGSOFT Symp. Found. Softw. Eng. (ESEC-FSE)*, 2011, pp. 416–419.
- [16] R. J. Grissom and J. J. Kim, *Effect Sizes for Research: A Broad Practical Approach*. Washington, DC, USA: Lawrence Erlbaum Associates Publishers, 2005.
- [17] S. Holm, "A simple sequentially rejective Bonferroni test procedure," *Scand. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.
- [18] Y. Hong, C. Tantithamthavorn, P. Thongtanunam, and A. Aleti, "CommentFinder: A simpler, faster, more accurate code review comments recommendation," in *Proc. 30th ACM Joint Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng. (ESEC-FSE)*, 2022, pp. 507–519.
- [19] J. Jiang, D. Lo, J. Zheng, X. Xia, Y. Yang, and L. Zhang, "Who should make decision on this pull request? Analyzing time-decaying relationships and file similarities for integrator prediction," *J. Syst. Softw.*, vol. 154, no. C, pp. 196–210, 2019.
- [20] J. Jiang, Y. Yang, J. He, X. Blanc, and L. Zhang, "Who should comment on this pull request? Analyzing attributes for more accurate commenter recommendation in pull-based development," *Inf. Softw. Technol.*, vol. 84, no. C, pp. 48–62, Apr. 2017, doi: 10.1016/j.infsof.2016.10.006.
- [21] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proc. 8th Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 66–71.
- [22] L. Li et al., "AUGER: Automatically generating review comments with pre-training models," in *Proc. 30th ACM Joint Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng. (ESEC/FSE)*, 2022, pp. 1009–1021.
- [23] Z. Li et al., "Automating code review activities by large-scale pre-training," in *Proc. 30th ACM Joint Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng. (ESEC/FSE)*, 2022, pp. 1035–1047.
- [24] M. V. Mäntylä and C. Lassenius, "What types of defects are really discovered in code reviews?," *IEEE Trans. Softw. Eng.*, vol. 35, no. 3, pp. 430–448, May/Jun. 2009.
- [25] A. Mastropaolo, L. Pascarella, and G. Bavota, "Using deep learning to generate complete log statements," in *Proc. 44th IEEE/ACM Int. Conf. Softw. Eng. (ICSE)*, 2022, pp. 2279–2290.
- [26] A. Mastropaolo et al., "On the robustness of code generation techniques: An empirical study on GitHub Copilot," in *Proc. 45th IEEE/ACM Int. Conf. Softw. Eng. (ICSE)*, pp. 2149–2160.
- [27] S. McIntosh, Y. Kamei, B. Adams, and A. E. Hassan, "The impact of code review coverage and code review participation on software quality: A case study of the Qt, VTK, and ITK projects," in *Proc. 11th IEEE/ACM Work. Conf. Mining Softw. Repositories (MSR)*, 2014, pp. 192–201.
- [28] E. Mirsaedi and P. C. Rigby, "Mitigating turnover with code review recommendation: Balancing expertise, workload, and knowledge distribution," in *Proc. 42nd ACM/IEEE Int. Conf. Softw. Eng. (ICSE)*, 2020, pp. 1183–1195.
- [29] R. Morales, S. McIntosh, and F. Khomh, "Do code review practices impact design quality? A case study of the Qt, VTK, and ITK projects," in *Proc. 22nd Int. Conf. Softw. Anal., Evolution, Reeng. (SANER)*, 2015, pp. 171–180.
- [30] A. Ouni, R. G. Kula, and K. Inoue, "Search-based peer reviewers recommendation in modern code review," in *Proc. 32nd IEEE Int. Conf. Softw. Maintenance Evolution (ICSME)*, 2016, pp. 367–377.
- [31] C. Pacheco and M. D. Ernst, "RANDOOP: Feedback-directed random testing for Java," in *Proc. ACM/SIGPLAN Int. Symp. New Ideas, New Paradigms, Reflections Program. Softw. (OOPSLA)*, 2007, pp. 815–816.
- [32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2002, pp. 311–318.
- [33] L. Pascarella, D. Spadini, F. Palomba, M. Bruntink, and A. Bacchelli, "Information needs in contemporary code review," in *Proc. ACM Hum.-Comput. Interact.*, 2018, vol. 2, no. CSCW, Art. no. 135, pp. 1–27.
- [34] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020, Art. no. 140.
- [35] M. M. Rahman, C. K. Roy, and J. A. Collins, "CoRRECT: Code reviewer recommendation in GitHub based on cross-project and technology experience," in *Proc. 38th Int. Conf. Softw. Eng. (ICSE)*, 2016, pp. 222–231.
- [36] P. C. Rigby and C. Bird, "Convergent contemporary software peer review practices," in *Proc. 21st ACM/SIGSOFT Joint Meeting Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng. (ESEC-FSE)*, 2013, pp. 202–212.
- [37] P. C. Rigby, D. M. Germán, L. L. E. Cowen, and M. D. Storey, "Peer review on open-source software projects: Parameters, statistical models, and theory," *ACM Trans. Softw. Eng. Methodol.*, vol. 23, no. 4, pp. 35:1–35:33, 2014.
- [38] B. Rosner, *Fundamentals of Biostatistics*, 7th ed. Boston, MA, USA: Brooks/Cole, 2011.
- [39] C. Sadowski, E. Söderberg, L. Church, M. Sipko, and A. Bacchelli, "Modern code review: A case study at Google," in *Proc. 40th Int. Conf. Softw. Eng.: Softw. Eng. Pract. (ICSE-SEIP)*, 2018, pp. 181–190.
- [40] S. Shi, M. Li, D. Lo, F. Thung, and X. Huo, "Automatic code review by learning the revision of source code," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 4910–4917.
- [41] A. Strand, M. Gunnarson, R. Britto, and M. Usman, "Using a context-aware approach to recommend code reviewers: Findings from an industrial case study," in *Proc. 42nd Int. Conf. Softw. Eng., Softw. Eng. Pract. (ICSE-SEIP)*, 2020, pp. 1–10.
- [42] P. Thongtanunam, C. Pornprasit, and C. Tantithamthavorn, "AutoTransform: Automated code transformation to support modern code review process," in *Proc. IEEE/ACM 44th Int. Conf. Softw. Eng. (ICSE)*, 2022, pp. 237–248.
- [43] P. Thongtanunam, C. Tantithamthavorn, R. G. Kula, N. Yoshida, H. Iida, and K. Matsumoto, "Who should review my code? A file location-based code-reviewer recommendation approach for modern code review," in *Proc. 22nd IEEE Int. Conf. Softw. Anal., Evolution, Reeng. (SANER)*, 2015, pp. 141–150.
- [44] F. Tian and C. Treude, "Adding context to source code representations for deep learning," in *Proc. IEEE Int. Conf. Softw. Maintenance Evolution (ICSME)*, 2022, pp. 374–378.
- [45] N. Tsantalis, A. Ketkar, and D. Dig, "RefactoringMiner 2.0," *IEEE Trans. Softw. Eng.*, vol. 48, no. 3, pp. 930–950, Mar. 2022.

- [46] M. Tufano, J. Pantiuchina, C. Watson, G. Bavota, and D. Poshyvanyk, "On learning meaningful code changes via neural machine translation," in *Proc. 41st IEEE/ACM Int. Conf. Softw. Eng. (ICSE)*, 2019, pp. 25–36.
- [47] M. Tufano, C. Watson, G. Bavota, M. Di Penta, M. White, and D. Poshyvanyk, "An empirical study on learning bug-fixing patches in the wild via neural machine translation," *ACM Trans. Softw. Eng. Methodol.*, vol. 28, no. 4, pp. 19:1–19:29, 2019.
- [48] R. Tufano, S. Masiero, A. Mastropaolo, L. Pascarella, D. Poshyvanyk, and G. Bavota, "Using pre-trained models to boost code review automation," in *Proc. 44th IEEE/ACM Int. Conf. Softw. Eng. (ICSE)*, 2022, pp. 2291–2302.
- [49] R. Tufano, L. Pascarella, M. Tufano, D. Poshyvanyk, and G. Bavota, "Towards automating code review activities," in *Proc. 43rd IEEE/ACM Int. Conf. Softw. Eng. (ICSE)*, 2021, pp. 163–174.
- [50] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24824–24837.
- [51] F. Wilcoxon, "Individual comparisons by ranking methods," *Int. Biometric Soc.*, vol. 1, no. 6, pp. 80–83, 1945.
- [52] X. Xia, D. Lo, X. Wang, and X. Yang, "Who should review this change?: Putting text and file location analyses together for more accurate recommendations," in *Proc. 31th IEEE Int. Conf. Softw. Maintenance Evolution (ICSME)*, 2015, pp. 261–270.
- [53] Z. Xia, H. Sun, J. Jiang, X. Wang, and X. Liu, "A hybrid approach to code reviewer recommendation with collaborative filtering," in *Proc. 6th Int. Workshop Softw. Mining (SoftwareMining)*, 2017, pp. 24–31.
- [54] H. Ying, L. Chen, T. Liang, and J. Wu, "EARec: Leveraging expertise and authority for pull-request reviewer recommendation in GitHub," in *Proc. 3rd Int. Workshop CrowdSourcing Softw. Eng. (CSI-SE@ICSE)*, 2016, pp. 29–35.
- [55] S. Yu, T. Wang, and J. Wang, "Data augmentation by program transformation," *J. Syst. Softw.*, vol. 190, 2022, Art. no. 111304.
- [56] Y. Yu, H. Wang, G. Yin, and T. Wang, "Reviewer recommendation for pull-requests in GitHub: What can we learn from code review and bug assignment?," *Inf. Softw. Technol.*, vol. 74, no. C, pp. 204–218, Jun. 2016.
- [57] M. B. Zanjani, H. Kagdi, and C. Bird, "Automatically recommending peer reviewers in modern code review," *IEEE Trans. Softw. Eng.*, vol. 42, no. 6, pp. 530–543, Jun. 2016.



Rosalia Tufano (Student Member, IEEE) received the M.Sc. degree in applied mathematics from Università degli Studi di Napoli Federico II, Italy, in March 2019. She is working toward the Ph.D. degree with the Faculty of Informatics at the Università della Svizzera Italiana (USI), Switzerland, and part of the Software Analytics Research Team (SEART). Her research interests mainly include the study and the application of machine learning techniques to support code-related tasks. More information available at:

<https://www.inf.usi.ch/phd/tufanr/>.



Ozren Dabić is a Research Assistant and a Software Engineer with the Software Analytics Research Team (SEART) at the Software Institute, Lugano. As a Full Stack Developer, his work primarily revolves around creating software solutions and web platforms for academia, intended to streamline the research process. He also participates in research focusing on the usage of deep learning models to automate various software engineering tasks.



Antonio Mastropaolo (Student Member, IEEE) received the M.Sc. degree in software system security from the Università degli studi del Molise, Italy, in July 2020. He is working toward the Ph.D. degree with the Faculty of Informatics at the Università della Svizzera Italiana (USI), Switzerland, where he is part of the Software Institute. His research interests include the study and the application of deep-learning techniques to foster code-related tasks. More information available at: <https://antoniomastropaolo.com>.



Matteo Ciniselli (Graduate Student Member, IEEE) received the M.Sc. degree in mathematical engineering from Politecnico di Milano, Italy, in April 2015. He is working toward the Ph.D. degree with the Faculty of Informatics at the Università della Svizzera Italiana (USI), Switzerland. His research interests mainly include the study of deep-learning models to improve the performances of code-related tasks. More information available at: <https://www.inf.usi.ch/phd/cinism/>.



Gabriele Bavota (Member, IEEE) received the Ph.D. degree in computer science from the University of Salerno, Italy, in 2013. He is an Associate Professor with the Faculty of Informatics at the Università della Svizzera Italiana (USI), Switzerland, where he is part of the Software Institute and he leads the SEART research group. His research interests include software maintenance and evolution, code quality, mining software repositories, and empirical software engineering. On these topics, he authored over 150 papers appeared in international journals and conferences. He has received four ACM Sigsoft Distinguished Paper awards at the three top software engineering conferences: ASE 2013 and 2017, ESEC-FSE 2015, and ICSE 2015. He also received the best/distinguished paper award at SCAM 2012, ICSME 2018, MSR 2019, and ICPC 2020. He was the recipient of the 2018 ACM Sigsoft Early Career Researcher Award for outstanding contributions in the area of software engineering as an early career investigator and the principal investigator of the DEVINTA ERC project. More information is available at: <https://www.inf.usi.ch/faculty/bavota/>.