

LLM-Based Detection of Tangled Code Changes for Higher-Quality Method-Level Bug Datasets

Md Nahidul Islam Opu
SQM Research Lab
Computer Science
University of Manitoba
Winnipeg, Canada

Shaowei Wang
Mamba Lab
Computer Science
University of Manitoba
Winnipeg, Canada

Shaiful Chowdhury
SQM Research Lab
Computer Science
University of Manitoba
Winnipeg, Canada

Abstract

Tangled code changes, commits that conflate unrelated modifications such as bug fixes, refactorings, and enhancements, introduce significant noise into bug datasets and adversely affect the performance of bug prediction models. Addressing this issue at a fine-grained, method-level granularity remains unexplored. This is critical to address, as recent bug prediction models, driven by practitioner demand, are increasingly focusing on finer granularity rather than traditional class- or file-level predictions. This study investigates the utility of Large Language Models (LLMs) for detecting tangled code changes by leveraging both commit messages and method-level code diffs. We formulate the problem as a binary classification task and evaluate multiple prompting strategies, including zero-shot, few-shot, and chain-of-thought prompting, using state-of-the-art proprietary LLMs such as GPT-5 and Gemini-2.0-Flash, and open-source models such as GPT-OSS-120B and CodeBERT.

Our results demonstrate that combining commit messages with code diffs significantly enhances model performance, with the combined few-shot and chain-of-thought prompting achieving an F1-score of 0.883. Additionally, we explore machine learning models trained on LLM-generated embeddings, where a multi-layer perceptron classifier achieves superior performance (F1-score: 0.906, MCC: 0.807). Applying our approach to 49 open-source projects improves the distributional separability of code metrics between buggy and non-buggy methods, demonstrating the promise of LLMs for method-level commit untangling and potentially contributing to improving the accuracy of future bug prediction models.

Keywords

Tangled code changes, large language models, machine learning, code diffs, commit messages, prompting techniques

1 Introduction

Software maintenance often exceeds the cost of initial development [12], with bug detection and resolution alone consuming 50–70% of total development expenses [79]. Undetected bugs can result in cascading failures, reduced developer productivity, and increased technical debt [22, 73]. Consequently, the research community has extensively investigated bug prediction techniques to facilitate timely and targeted maintenance interventions [20, 27, 64, 66, 80]. Early bug prediction models focused on class- or file-level granularity [3, 7, 9, 28, 74, 86, 92]. However, these models are less helpful in practice, as only a small fraction of code in large units is typically defective [58, 64]. Line-level models offer finer granularity but are prone to false positives due to coincidental line similarity [20, 31, 68]. This has driven growing interest in method-level

bug prediction [20, 27, 34, 57, 58, 64, 71], which aligns more closely with real-world debugging practices [20, 31]. However, method-level bug prediction faces challenges including evaluation bias and noisy data [20, 64], and progress is constrained by the scarcity of clean, labeled datasets [20, 64].

A key obstacle in building clean bug datasets is the presence of *noise* in training data [11], often caused by *tangled code changes* [20, 36, 63]. While Version Control Systems like Git are intended for atomic commits, developers frequently combine unrelated tasks, such as bug fixes, refactorings, and enhancements, into a single commit [24, 37, 43], violating best practices and introducing tangled changes. Researchers typically use commit messages to identify bug-fix commits, assuming all associated modifications are relevant. However, this assumption fails in the presence of tangled commits, leading to false positives in labels, which degrades dataset quality and impairs bug prediction performance [35, 36].

To mitigate the impact of tangled changes, several prior works have proposed various untangling techniques ranging from heuristic-based program slicing [59], lexical and data-flow based models [63], and graph-based learning using dependency and name flows [48, 83]. While effective at the file or statement level, none of these methods have addressed the problem at the method-level granularity, leaving a gap that presents a unique opportunity for exploration. Method-level untangling cannot be achieved using file-level approaches, as a single file may contain multiple methods, only some of which are bug-related. Labeling all methods in a file as buggy is inappropriate and would produce numerous false positives. On the other hand, although statements can be aggregated to form method-level representations, determining whether an individual statement contributes to a bug fix requires understanding its broader context, including surrounding code, method semantics, and commit intent [72, 76]. For developers and models alike, the relevant context may span from within the method to the entire file or even across files [72]. For instance, in the *refill* function of *FixedIntervalRateLimiter.java* from *Apache HBase Commit 84a5039*¹, the function signature modification appears to be a bug fix when viewed in isolation, as removing an argument typically alters the function’s logic; however, examining the full method reveals it is actually a refactoring change. Thus, simply applying a statement-level model repeatedly at the method-level would require redundant context analysis for each modified line, introducing unnecessary computational overhead. Therefore, distinct representations at different granularity are needed to effectively model bug-fix behavior while maintaining contextual relevance.

¹<https://github.com/apache/hbase/commit/84a50393ee56d09abb68f54b44b64f5279bd33c9>

Given that, this study focuses on detecting tangled changes at the method-level by framing the task as a binary classification problem: determining whether a method-level code change is related to a bug fix (*Buggy*) or not (*NotBuggy*) based on commit messages and code diffs. To tackle this, we explore the use of *Large Language Models* (LLMs), which have demonstrated strong capabilities in analyzing natural language and source code [18, 60]. In contrast to prior approaches in code analysis, classification, and untangling that rely on language-specific algorithms [32, 36] or complex models requiring extensive training and feature engineering [88], LLM-based methods enable faster system development through their language-agnostic design. LLMs also provide advantages such as promptability, flexibility in input handling, and the ability to leverage contextual signals from both code and commit metadata.

We evaluate several prompting strategies using both proprietary and open-source LLMs. Additionally, we develop machine learning models using LLM generated embeddings. Finally, we evaluate the impact of the LLM-based technique by measuring improvements in the distributional separability of code metrics between *Buggy* and *NotBuggy* methods through statistical analysis. In general, the contribution of the paper is centered around four research questions.

RQ1: Can zero-shot LLMs detect tangled code changes by using code diff and commit message?

We designed a classification-based approach that uses LLMs to detect tangled changes from method-level code diffs, using zero-shot prompting, with or without the corresponding commit messages. Through experiments on our curated gold dataset, we show that including the commit message significantly improves the performance across multiple LLM variants, achieving the highest F1-score of 0.879 using *gpt-5*.

RQ2: How do different prompting techniques influence the effectiveness of detecting tangled code changes?

We investigate the impact of advanced prompting strategies, such as few-shot, chain-of-thought, and a hybrid combination. Our analysis shows that LLMs can reason more accurately about change semantics when guided with structured prompts and examples. Among these, chain-of-thought + few-shot prompting with *gpt-4o* achieves the most balanced performance (F1-score: 0.883) and demonstrates superior capability in handling complex inputs.

RQ3: How well can embedding-based machine learning models detect tangled code changes?

We utilized LLMs to generate embeddings from commit messages and code diffs, which served as input for embedding-based classifiers. One of the models exhibits strong predictive capabilities, outperforming the results obtained in RQ1 and RQ2. Specifically, a Multi-layer Perceptron classifier achieved the highest F1-score of 0.906, demonstrating the effectiveness of embedding-based representations for detecting tangled changes.

RQ4: What is the potential impact of LLM-based untangling on future method-level bug prediction models?

Using a *Less-Noisy* dataset, created by filtering out noisy samples through our LLM-based approach, we observe that the distributional differences in various code metrics between *Buggy* and *NotBuggy* methods are significantly more than in the original noisy dataset. This finding shows promise for future machine learning

(ML) models for bug prediction, as many of them rely on these code metrics to predict whether a method is bug-prone or not.

To enable replication, we share our data and code publicly.²

2 Related Works

This section reviews prior work on tangled code changes, their prevalence, impact, and untangling techniques, as well as LLM-based approaches for analyzing and classifying text and code. Together, these studies motivate our research.

2.1 Tangled Changes

Herzig and Zeller [37] first used the term *Tangled Changes* and reported that up to 15% of bug-fixing commits across several open-source Java projects contained tangled changes. Subsequently, Tao and Kim [75] later observed tangling in up to 29% of revisions. Herbold et al. [35] confirmed its prevalence and impact, showing that only 22–38% of lines in bug-fix commits actually fix bugs. Kochhar et al. [44] further found that 28% of files in bug-fix commits contain no bugs.

The prevalence of tangled changes greatly affects the historical analysis of source code. By labeling all the modified methods in a bug-fixing commit as buggy methods, it harms the bug prediction models significantly [20, 36]. Consequently, researchers focused on detecting and untangling commits [32, 37, 59, 63, 75].

Early techniques relied on static analysis and heuristics. Herzig et al. [37] introduced confidence voters based on file distance, change coupling, call graphs, and data dependencies to estimate task co-membership. Tao and Kim [75] applied program slicing and pattern matching to cluster semantically related edits, achieving 69% agreement with manual decompositions. Guo and Song [32] proposed CHGCUTTER, an interactive method using control and data dependencies to decompose composite changes while preserving syntactic correctness. Muylaert and De Roover [59] applied program slicing on abstract syntax trees to untangle composite commits by grouping fine-grained code changes according to their dependence graph slices. Pärtachi et al. [63] introduced Flexeme, which overlays name flows on program dependency graphs using a δ -NFG structure to better capture lexical cues. Using Agglomerative clustering for commit untangling they developed a tool Heddle, which achieved improved clustering accuracy (F1-score 0.81) and runtime efficiency.

Graph-based modeling has emerged as a powerful approach for capturing complex code relationships. Shen et al. [70] proposed SmartCommit, representing diff hunks as graph nodes enriched with semantic links (hard, soft, refactoring, cosmetic), and achieved a decomposition accuracy between 71–84%. Chen et al. [16] and Xu et al. [83] enhanced this approach by incorporating node attributes such as token content and control/data flow, achieving 7.8% and 8.2% higher accuracy than Flexeme, respectively.

Fan et al. [25] introduced a Heterogeneous Directed Graph Neural Network designed to capture semantic dependencies without relying on explicit code links. By leveraging hierarchical graphs at the entity and statement levels, they achieved substantial untangling improvements, 25% for C# and 19.2% for Java, outperforming Flexeme and SmartCommit without sacrificing time efficiency.

²<https://github.com/SQMLab/Tangled>

Recent work increasingly applies machine learning to learn untangling patterns from data. UTANGO [48] uses a Graph Convolutional Network to learn contextual embeddings of code changes, incorporating cloned code and surrounding context. Framed as supervised clustering, it outperforms Flexeme by 9.9% accuracy. Dias et al. [24] and Liu et al. [50] explored fine-grained tracking and clustering based on developer interactions and commit metadata to achieve high temporal and contextual resolution.

Although these methods have shown success at the file- or statement-level, the method-level remains unexplored, despite being preferred by practitioners and researchers in order to create a clean and noise-free method-level bug dataset [20, 64]. Motivated by this gap, we investigate whether tangled changes can be detected from method-level diffs and their corresponding commit messages.

2.2 LLMs for Code Analysis & Classification

LLMs have shown strong performance in natural language understanding and source code analysis [13, 90]. Trained on large corpora of text and code, they generalize well to unseen inputs and often perform effectively in zero-shot settings [45]. This versatility has fueled growing interest in using LLMs for text classification and code analysis [2, 26]. Building on these capabilities, our study explores LLMs for detecting tangled code changes by jointly analyzing commit messages and code diffs.

Prompt design is critical to optimizing LLM performance. Recent advances in prompt engineering have introduced zero-shot, few-shot, and chain-of-thought prompting [13, 81, 82]. Zero-shot learning uses only task instructions, while few-shot prompting includes exemplar input-output pairs to guide responses, yielding improved performance across tasks [13]. Chain-of-thought prompting enhances reasoning by encouraging step-wise problem decomposition [82]. Motivated by their empirical success, we systematically evaluate these techniques for detecting tangled code changes.

LLMs are built on Transformer architectures, consisting of encoder and decoder modules [77]. Encoders convert input text into high-dimensional embeddings that capture semantic relationships, making them effective for downstream classification [42, 65]. We extend this utility to software engineering by generating embeddings from commit messages and code diffs, and training machine learning models to classify them as tangled or untangled.

LLM embeddings are also used in commit classification tasks [87, 88], which classifies commits into maintenance activity types. CodeBERT [26] has been used in COLARE [88] to classify commits into corrective, adaptive, and perfective categories by integrating hunk-level code representations with commit messages and file features. However, such models are built for commit-level classification which does not properly align with our method-level objectives.

In summary, this study is motivated by three factors: (1) the lack of fine-grained research on the tangled changes at the method level, especially for bug prediction [20, 36]; (2) the proven capabilities of LLMs in analyzing textual and code artifacts via advanced prompting [13, 81, 82, 90]; and (3) the semantic richness of LLM-generated embeddings for classification [42, 65]. *To the best of our knowledge, this is the first study to evaluate LLMs for detecting tangled code changes within bug-fix commits at the method level.*

3 Methodology

Figure 1 shows the overview of our approach that we use to evaluate LLMs in distinguishing *Buggy* diff and *NotBuggy* diff. In this section, the components of the methodology are described step by step.

3.1 Dataset

To assess whether LLMs can be used to untangle real bug fix changes from other changes, a dataset is required that includes both genuine bug fix code diffs and non-bug-fix code diffs. We provide the model with a code diff with or without the commit message to evaluate its ability to distinguish between bug-fix and non-bug-fix changes.

Fortunately, an existing dataset by Chowdhury et al. [20] contains 774,051 Java methods from 49 open source projects, where each method is associated with its complete change history and is enriched with metadata such as commit messages, timestamps, authorship, and change type. The dataset also includes code diffs for all historical changes, labels indicating whether each change is related to a bug fix or not, and information on the number of methods modified within the same commit.

However, the dataset contains noise. In some cases, methods involved in a bug fix commit may not contribute to the actual fix, representing instances of tangled changes. We use this dataset to construct a curated gold dataset by removing such noise. The construction process involves a two-step sampling strategy, combining automated filtering and manual labeling, to ensure high quality and accurate representation of tangled change scenarios.

Automated Approach. The dataset provided by Chowdhury et al. [20] includes the total number of methods modified in each commit. They recommended that if a method is the only one modified in a bug fix commit, it can be confidently considered a bug-related change. We classify such code diffs as *Buggy* without ambiguity. Using this criterion, we collected 730 *Buggy* diffs. They also suggested that a method can be considered *NotBuggy* if it has never appeared in a bug fix commit throughout its lifetime. Based on this guideline and leveraging the large pool of eligible *NotBuggy* methods, we randomly selected 730 samples that have no history of being part of any bug fix commit. During this process, three duplicate diffs, likely caused by duplicated code across codebases, were identified and removed, resulting in 727 unique *NotBuggy* instances. Thus, the *Automated* step of our gold dataset construction produced a total of 1,457 method-level examples. We intentionally limited the number of *NotBuggy* samples, as querying proprietary LLMs incurs significant computational and financial costs. Including all such samples would have significantly increased the overall cost.

Manual Approach. While the *Automated* approach ensured high confidence in labeling, it did not capture the complexity of real-world tangled change scenarios, where multiple unrelated modifications often occur within a single commit. To address this limitation, we introduced a second step focused on incorporating tangled examples. In this step, we extracted additional method-level diffs from bug-fix commits that modified multiple methods. The first author, a graduate student with over two years of industry experience in software engineering, manually labeled each method-level change to determine whether it was related to a bug fix (*Buggy*) or was incidental, such as refactorings or unrelated enhancements (*NotBuggy*). Each method was reviewed in conjunction with its associated commit message to evaluate semantic alignment with

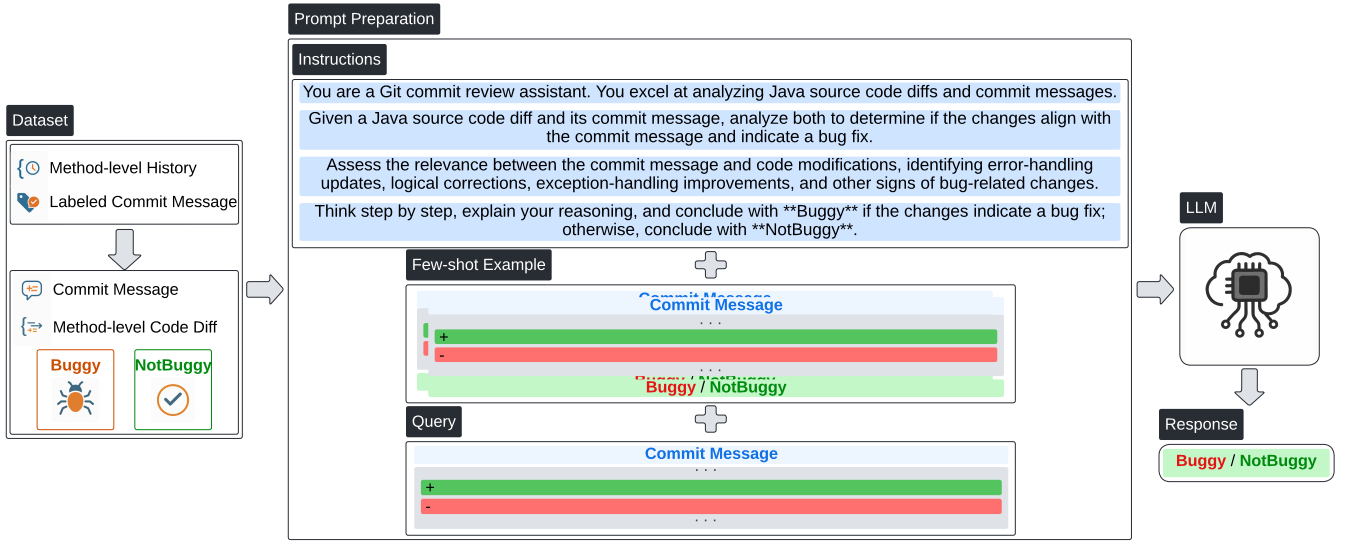


Figure 1: The methodology for RQ1 and RQ2 consists of three parts: (a) dataset, (b) prompt preparation, and (c) LLM Query. The prompt setup for few-shot + chain-of-thought prompting consists of four instruction parts: (i) persona, (ii) task description, (iii) behavioral guidance, and (iv) output formatting. For Diff-only detection of RQ1, references to the commit message are excluded from everywhere. Other experiments use the full instructions, with the output formatting modified to require a single-word answer in non-chain-of-thought setups. Few-shot examples are included only in few-shot experiments.

bug-fixing intent. The evaluation emphasized structural corrections, error handling logic, and domain-specific repairs. To promote annotation consistency and minimize errors, we developed a user interface that allowed side-by-side inspection of code diffs and their corresponding commit metadata. Using this *manual* approach, we identified 166 tangled *Buggy* methods and 141 tangled *NotBuggy* methods from 450 randomly selected commits in the dataset of Chowdhury et al. [20].

Combining all steps, the gold dataset consists of 1,764 method-level change instances, each labeled as either *Buggy* or *NotBuggy*, based on their actual relevance to bug fixing activities. To evaluate labeling reliability, an independent validation was conducted by another graduate student in software engineering, who randomly picked 100 samples from the manually labeled portion of the dataset and labeled them independently. Some disagreements arose, mainly due to differing interpretations of what constitutes a bug. Specifically, the first labeler marked some quality-related fixes as *NotBuggy*, while the second considered them *Buggy*. Following a discussion with the last author, quality-fix-related changes were finalized as *NotBuggy*. The inter-rater agreement, measured using Cohen’s Kappa [23], was 0.82, indicating strong consistency and confirming the reliability of the labeling process. Consequently, this dataset provides a robust benchmark for both training and evaluating our detection approaches, while also offering a valuable foundation for future research in this domain.

3.2 Prompt Preparation

The design of our prompt structure draws on recent advancements in prompt engineering that demonstrate the effectiveness of decomposing complex instructions into clearly scoped components [13, 82, 91]. As prior work has shown, LLMs are highly sensitive to prompt formulation, and performance can vary substantially

based on how the task, instructions, and examples are framed [49]. To maximize reliability and adaptability, we designed a modular prompt template consisting of three core components, aligned with prior studies [4, 29], as illustrated in Figure 1.

Instructions. This component defines the LLM’s role and behavior. It is composed of four parts: (i) a role-based persona (e.g., “You are a Git commit review assistant”), which helps constrain the model’s responses by grounding its behavior and improves performance [39]; (ii) a high-level task description [4]; (iii) step-by-step behavioral guidance to structure the model’s reasoning [82]; and (iv) explicit output formatting requirements to reduce ambiguity and improve parsability [51].

Few-shot Examples. Few-shot prompting has been shown to significantly improve performance by enabling in-context learning, especially in code and software engineering tasks [90]. For few-shot examples, as shown in Figure 2, we selected a representative bug-fix commit containing one diff genuinely related to a bug fix and another unrelated change, a real instance of tangled code changes. For the negative example (*NotBuggy*), we selected the code diff that merely replaced constant values with equivalent boolean literals, without altering the program’s logic or behavior. In contrast, the other code diff (*Buggy*) addressed an issue in the byte-to-hexadecimal conversion logic by correctly handling signed bytes using bit masking, rectifying a functional error. These two fixed examples were consistently utilized to prime the LLMs in all experiments employing few-shot learning. These examples were selected to represent contrasting scenarios, one clearly indicative of a bug-fix and the other representing a refactor or performance-related change that does not qualify as a bug-fix, although both of the examples come from the same commit.

LLMs often exhibit significant performance degradation as prompt length increases, due to limitations in handling extended context [47,

```

Fix bug in InetAddresses and performance problem in InternetDomainName (changed from regex to CharMatcher). This is intended as the last code
update before release 06.
@@ -1,18 +1,18 @@
- private static boolean validateSyntax(List<String> parts) {
+ private static boolean validateSyntax(List<String> parts) {
    final int lastIndex = parts.size() - 1;
    // Validate the last part specially, as it has
    different syntax rules.
-    if (!validatePart(parts.get(lastIndex), FINAL_PART)) {
+    if (!validatePart(parts.get(lastIndex), true)) {
        return false;
    }
    for (int i = 0; i < lastIndex; i++) {
        String part = parts.get(i);
-        if (!validatePart(part, NORMAL_PART)) {
+        if (!validatePart(part, false)) {
            return false;
        }
    }
    return true;
}

@@ -1,12 +1,12 @@
- private static String convertDottedQuadToHex(String ipString) {
+ private static String convertDottedQuadToHex(String ipString) {
    int lastColon = ipString.lastIndexOf(':');
    String initialPart = ipString.substring(0, lastColon + 1);
    String dottedQuad = ipString.substring(lastColon + 1);
    byte[] quad = textToNumericFormatV4(dottedQuad);
    if (quad == null) {
        return null;
    }
    String penultimate = Integer.toHexString((quad[0] << 8) |
(quad[1] & 0xff));
    String ultimate = Integer.toHexString((quad[2] << 8) | (quad[3] &
0xff));
+    String penultimate = Integer.toHexString(((quad[0] & 0xff) << 8)
+ (quad[1] & 0xff));
+    String ultimate = Integer.toHexString(((quad[2] & 0xff) << 8) |
(quad[3] & 0xff));
    return initialPart + penultimate + ":" + ultimate;
}

```

Figure 2: Commit message and two methods modified in the same commit: the message indicates a bug fix, but only the method on the right contains the bug-fix change, while the method on the left is a non-bug-related tangled change. This was used as few-shot examples in our prompts.

89]. Additionally, longer prompts incur higher computational and monetary costs, making token efficiency a practical concern. Since method-level diffs can vary greatly in length and the input query, over which we have no control, is the primary driver of prompt size, we deliberately restricted the prompt to a single example per class (Buggy and NotBuggy). This prompting approach helps maintain prompt lengths within manageable token limits across different model families. By limiting the number of in-context examples, we strike a balance between performance and resource usage. This strategy aligns with established practices in few-shot prompting [13], where concise and diverse instances are used to anchor the model’s reasoning prior to inference.

Query. This includes the method-level code diff and its corresponding commit message.

Each research question required slight variations in prompt configuration, e.g., enabling or disabling few-shot examples, toggling chain-of-thought reasoning, or isolating input modalities (diff-only vs. diff + message). All corresponding prompt templates are included in our replication package, located within the *prompts* directory.

3.3 Selecting and Querying LLMs

Selecting LLMs. For RQ1 and RQ2, we selected six LLMs: four proprietary, *GPT-4o-mini*, *GPT-4o*, *GPT-5*, and *Gemini-2.0-Flash* and two open-source, *GPT-OSS-20B* and *GPT-OSS-120B*. The proprietary LLMs were chosen based on their performance on recent code-related benchmarks [8, 14, 30, 62], robust API support, and cost-effectiveness for large-scale experimentation. *GPT-4o* serves as OpenAI’s full-scale model offering superior robustness over its smaller variant, *GPT-4o-mini* [14, 62], while *GPT-5* represents the latest generation of OpenAI’s LLMs. *Gemini-2.0-Flash*, from Google [30], delivers competitive performance with the GPT series and has demonstrated effectiveness in code-related tasks [8].

For open-source alternatives and to ensure reproducibility and accessibility, we choose *GPT-OSS-20B* (requiring 16 GB GPU memory) and *GPT-OSS-120B* (runnable on a single 80 GB GPU) [1]. Despite their moderate scale, these models deliver reasoning and code-understanding performance comparable to proprietary base-lines [1, 10], balancing efficiency and capability.

For RQ3, we employed four models for embedding generation: one proprietary and three open-source. The proprietary model, Gemini’s *text-embedding-004* offers high-quality embeddings optimized for classification tasks [46]. Open-source models, *CodeBERT* [26] and *GraphCodeBERT* [33] are widely adopted for code representation learning [52], while *EmbeddingGemma-300m*, a recent lightweight model from Google DeepMind, is designed for low-latency, high-throughput applications [78]. The models used in RQ1 and RQ2 were not included in RQ3 because they are decoder-only generative models optimized for text generation, whereas RQ3 required encoder-based architectures specialized for producing fixed-length embeddings suitable for classification tasks.

In summary, proprietary models were selected for their state-of-the-art performance across benchmarks, including code-related tasks [5, 8], while open-source models were selected for their competitive accuracy and efficiency relative to larger open-source alternatives.

Querying LLMs. The three parts of the prompt (instruction + examples + query) are concatenated and formatted according to the input constraints of the respective LLM. The final query prompt is submitted to the LLM, which classifies the change as either *Buggy* or *NotBuggy*. Outputs are reviewed and filtered automatically, if required, to ensure alignment with evaluation criteria.

3.4 Evaluation Metrics

We evaluate model performance using standard classification metrics, including accuracy, precision, recall, and F1-score. Additionally, we use the Matthews Correlation Coefficient (MCC), a robust metric that considers all four outcomes: true positives, true negatives, false positives, and false negatives. Unlike accuracy, MCC provides a more balanced measure. An MCC of 1 indicates perfect prediction, 0 indicates random guessing, and -1 reflects total disagreement between predictions and actual labels.

3.5 Experimental Setup

All experiments involving open-source LLMs were conducted on a High-Performance Computing (HPC) cluster equipped with Intel 8570 CPUs (2.1 GHz) and multiple NVIDIA H100 GPUs. The number

of CPUs, GPUs and the memory allocation varied by model, depending on resource availability and scheduling within the cluster.

4 Approach, Analysis & Results

This section answers the four research questions and outlines the specific methodologies used to investigate each one.

4.1 RQ1: Performance of zero-shot LLMs using code diff and commit message

Both the commit message and the corresponding code diff contain information about the nature of the code changes. While the diff captures the structural modifications to the source code, the commit message provides a natural language description of the developer’s intent. LLMs offer a powerful mechanism to detect semantic signals indicative of bug-fixing behavior across these modalities. By leveraging their ability to reason over structured and unstructured data (code + natural language) [18, 60], LLMs offer a promising alternative to traditional rule-based or static analysis techniques. Furthermore, using prompt-based inference avoids the need for task-specific fine-tuning, making the approach adaptable across repositories and languages. However, the effectiveness of LLMs is sensitive to input length, as inference cost increases with the number of tokens in the prompt. Given that both code diffs and commit messages can vary significantly in length, it becomes essential to evaluate whether both are necessary for effective detection or whether the code diff alone suffices. This research question seeks to empirically determine whether commit messages provide additive value beyond the code diff for detecting tangled changes.

We investigate two scenarios using LLMs by designing tailored prompting strategies for the binary classification of method-level diffs. Since no examples are provided within the prompt, this setup qualifies as zero-shot prompting.

Diff-Only. The LLM is provided only with the code diff and instructed to assess whether the change represents a bug fix (Buggy) or not (NotBuggy). The prompt is adapted from Figure 1, with all references to commit messages and few-shot examples removed, and the instruction modified to require a single-word response.

Diff+Message. The prompt includes both the code diff and its commit message, instructing the model to assess their semantic alignment to identify bug-fixing intent. This prompt follows the *Diff-Only Detection* setting, except that all references to commit messages are retained as in Figure 1.

Table 1 compares the performance of the evaluated LLMs under two input settings: *Diff-only* and *Diff+Message*. In the *Diff-only* scenario, *gpt-5* achieves the best overall performance, yielding the highest F1-score (0.767) and MCC (0.490). When commit messages are included, all models show substantial improvements. The *gpt-5* again attains the highest F1-score (0.879), while *gpt-4o* provides the best accuracy (0.874), precision (0.868) and MCC (0.747). Although *gpt-5* achieves the top F1-score, its high precision-recall gap indicates less balanced performance, whereas *gpt-4o* offers the most consistent results across all metrics.

The *gemini-2.0-flash* and the *gpt-oss* models achieve high recall (> 0.9) but low precision (< 0.76) resulting in reduced F1-scores. The *gpt-oss* models perform remarkably well in *Diff+Message* scenario despite their smaller size narrowing the gap with proprietary

models (within 3–4% of *gpt-5*’s F1-score), likely benefiting from their unique dual-channel chat format [1], where an internal “analysis” channel generates intermediate reasoning (chain-of-thought) before producing the answer on “final” channel. Overall, these results demonstrate that integrating commit messages with diffs significantly improves performance, and that open-source LLMs are also effective despite their smaller sizes.

Table 1: Performance comparison of LLMs for Diff-only and Diff+Message. Metrics are calculated by targeting the buggy samples (pos_label = "Buggy"). For each scenario, the best value for each metric is highlighted in bold.

Scenario	LLM	Accuracy	Precision	Recall	F1-score	MCC
Diff-only	gpt-5	0.742	0.709	0.834	0.767	0.490
	gpt-4o-mini	0.704	0.660	0.860	0.747	0.426
	gpt-4o	0.691	0.654	0.831	0.732	0.395
	gemini-2.0-flash	0.692	0.650	0.853	0.738	0.402
	gpt-oss-20b	0.663	0.615	0.900	0.730	0.364
	gpt-oss-120b	0.667	0.619	0.900	0.733	0.372
Diff + Message	gpt-5	0.871	0.839	0.924	0.879	0.746
	gpt-4o-mini	0.822	0.764	0.941	0.843	0.661
	gpt-4o	0.874	0.868	0.886	0.877	0.747
	gemini-2.0-flash	0.796	0.717	0.988	0.831	0.639
	gpt-oss-20b	0.820	0.758	0.950	0.843	0.662
	gpt-oss-120b	0.827	0.766	0.948	0.847	0.671

Summary of RQ1: Both proprietary and open-source LLMs can effectively detect tangled code changes using code diffs, with or without commit messages, though combining both inputs yields better results than code diffs alone.

4.2 RQ2: Performance of different prompting techniques

The effectiveness of LLMs highly depends on the input prompt, and the success of different prompting techniques is proven in various studies [13, 81, 82]. This research question investigates how different prompting techniques, namely *few-shot learning*, *chain-of-thought prompting*, and a hybrid approach combining both, impact the classification performance of LLMs.

Few-shot. The prompt includes labeled examples of buggy and non-buggy changes to guide the model, while the instruction remains identical to the *Diff+Message* setting of RQ1.

Chain-of-thought. The model is instructed to explain its reasoning in the output. The exact instruction prompt is shown in Figure 1, and no examples are included.

Chain-of-thought + Few-shot. This setting combines the above two techniques by including labeled examples and requiring the model to explain its reasoning in the output.

Based on the findings from RQ1, we use the combination of commit message and code diff for RQ2. Additionally, we exclude *gpt-4o-mini* from further experiments, as *gpt-4o* significantly outperformed this small model. This decision also supports cost efficiency, as the prompting techniques used in RQ2 involve substantial token consumption in both input and output.

Table 2 presents the performance of different prompting strategies across LLMs. While *gpt-4o* with few-shot achieves the highest precision (0.903) but lower recall (0.833), other models exhibit the

opposite trend. In this setting, *gpt-5* attains the best overall performance with an F1-score of 0.884. For chain-of-thought prompting, performance generally decreases across models, though *gpt-5* again achieves the highest F1-score (0.872). When combining few-shot and chain-of-thought prompting, *gpt-4o* delivers the best results across all metrics except recall, which is highest for *gemini-2.0-flash*.

Table 2: Performance comparison of different prompting techniques. Metrics are calculated by targeting the buggy samples (pos_label = "Buggy").

Prompting Technique	LLM	Accuracy	Precision	Recall	F1-score	MCC
Few-shot	gpt-5	0.876	0.842	0.931	0.884	0.757
	gpt-4o	0.870	0.903	0.833	0.866	0.742
	gemini-2.0-flash	0.802	0.723	0.989	0.835	0.649
	gpt-oss-20b	0.837	0.775	0.958	0.857	0.694
	gpt-oss-120b	0.819	0.753	0.959	0.843	0.663
Chain of thought	gpt-5	0.863	0.829	0.920	0.872	0.730
	gpt-4o	0.827	0.770	0.940	0.846	0.669
	gemini-2.0-flash	0.765	0.688	0.984	0.810	0.587
	gpt-oss-20b	0.829	0.770	0.946	0.849	0.676
	gpt-oss-120b	0.829	0.765	0.956	0.850	0.679
Chain of thought + Few-shot	gpt-5	0.863	0.829	0.921	0.873	0.731
	gpt-4o	0.880	0.871	0.896	0.883	0.760
	gemini-2.0-flash	0.807	0.731	0.979	0.837	0.651
	gpt-oss-20b	0.844	0.795	0.932	0.858	0.697
	gpt-oss-120b	0.816	0.752	0.952	0.840	0.655

Among all prompting strategies, *gpt-5* with few-shot yields the top F1-score, while *gpt-4o* attains the highest accuracy, precision, and MCC with chain-of-thought+few-shot. Considering balanced performance across metrics, particularly the small gap between precision and recall (0.025), *gpt-4o* with chain-of-thought+few-shot offers the most balanced results, with an F1-score (0.883) nearly identical to the best (0.884). Open-source models also benefit consistently, though modestly, from advanced prompting; notably, *gpt-oss-20b* remains competitive, often matching or surpassing *gpt-oss-120b*.

A comparison between Table 1 and Table 2 reveals that few-shot prompting improves performance primarily for *gpt-5* and slightly for *gpt-oss-20b*, while other models show marginal or decreased performance. Notably, chain-of-thought alone results in poorer performance across most models. However, combining few-shot prompting with chain-of-thought yields the most balanced and consistent outcomes overall, as observed with *gpt-4o*. This trend aligns with prior research suggesting that few-shot prompting can lead models to overfit to example formats rather than the task itself, while zero-shot chain-of-thought prompting may amplify biases and degrade reasoning quality [45, 69]. Our findings thus corroborate earlier studies indicating that integrating few-shot and chain-of-thought generally outperforms zero-shot approaches [45].

4.2.1 Case analysis: an example of the effectiveness of chain-of-thought approach in a semantically ambiguous commit. To better understand why the chain of thought technique improves model performance in the few shot setting with *gpt-4o*, we analyze a representative example involving semantic ambiguity between the commit message and the code modifications from the *PMD*³ project.

³<https://github.com/pmd/pmd>

As shown in Table 3, the commit message clearly describes a bug fix involving two files, *PMD.java* and *ImmutableFieldRule.java*. In this commit⁴, multiple methods are modified, including *initializedInConstructor()* from *ImmutableFieldRule.java*, which only includes variable renaming (e.g., *occurrence* to *occ*) without affecting the program’s logic or behavior. No structural fix is present that addresses the reported issue. Despite this, both the zero-shot and few-shot models incorrectly classify the change as *Buggy*, influenced by the commit message. In contrast, the chain-of-thought technique, with or without few-shot prompting, correctly identifies the change as *NotBuggy*, consistent with the manually assigned label.

The reasoning produced by the chain-of-thought prompt illustrates a more grounded assessment. An important part of the reasoning states: ". . . However, in the provided diff, the only change is renaming the variable *occurrence* to *occ*, which is a simple refactor for readability or consistency purposes. This modification itself doesn’t address the logic or behavior of the code in terms of resolving the stated bug. The commit message also mentions a version number fix in *PMD.java*, but no relevant changes are shown in this diff. . ."

This case highlights the effectiveness of chain-of-thought in systematically analyzing complex commit messages that reference multiple issues. While such messages offer valuable context, they can also introduce ambiguity. The step-by-step reasoning decomposed the message, independently assessed each part against the code diff, and correctly identified the lack of structural or semantic alignment, ultimately concluding that the change was not a bug fix.

Summary of RQ2: Combining chain-of-thought reasoning with few-shot prompting yields the most balanced and consistent performance across models. While few-shot prompting alone benefits mainly *gpt-5*, and chain-of-thought alone often reduces accuracy, their combination enhances reasoning stability and mitigates overfitting. This hybrid approach also improves interpretability, enabling models to correctly handle semantically misleading commits through more context-aware reasoning.

4.3 RQ3: Performance of embedding-based machine learning models

This research question explores the viability of embedding-based ML models for detecting tangled code changes. Unlike prompt-based approaches that rely on in-context reasoning of LLMs, this method uses embeddings, vector representations of the input data from commit messages and code diffs, to train supervised classifiers. The key objective is to assess whether these embeddings, when used as features in ML models, can achieve competitive performance in detecting bug-related code changes. This experiment is motivated by the previous studies where embedding-based models performed well and even outperformed prompt-based techniques [42, 65].

The methodology employed for this research question diverges from that of the preceding research questions and the process illustrated in Figure 1. Specifically, this experiment omits the complex prompt engineering. Instead, only the core query, comprising the code diff and commit message, is retained and directly input to the LLM to generate fixed-length vector embeddings for downstream classification. The modified methodology is shown in Figure 3.

⁴<https://github.com/pmd/pmd/commit/a405d23dfb9e574e2b2ef23f1f45d548a738ed3b>

Table 3: Predictions on a semantically misleading commit using chain-of-thought + few-shot prompting with *gpt-4o*.

Commit Message	Diff (Excerpt)	Ground Truth	Prompting Method	Prediction
Fixed bug 1050173 – <i>ImmutableFieldRule</i> no longer reports false positives for static fields. Also fixed version number in PMD.java.	... - <i>NameOccurrence occurrence</i> = ... - <i>if (occurrence.isOnLeftHandSide())</i> { ... + <i>NameOccurrence occ</i> = ... + <i>if (occ.isOnLeftHandSide())</i> { ...	NotBuggy	Commit msg + diff (zero-shot)	Buggy
			Few-shot	Buggy
			Chain-of-thought	NotBuggy
			Chain-of-thought + Few-shot	NotBuggy

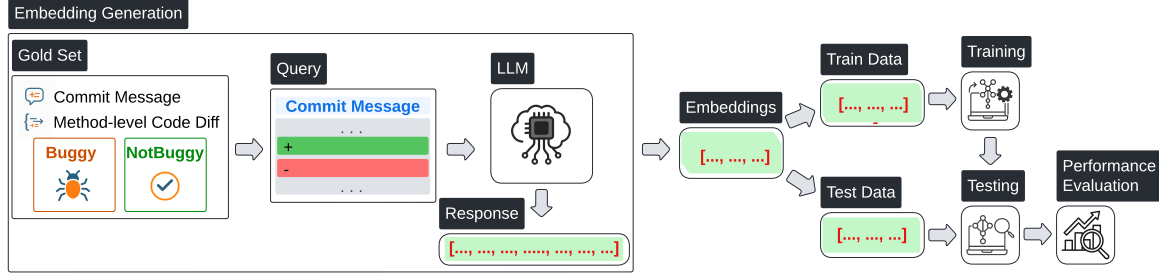


Figure 3: The methodology for RQ3 has two main parts: (i) embedding generations and (ii) building ML models.

To generate embeddings, Gemini’s *text-embedding-004* was employed with the *classification* task type. Due to resource constraints, this study focuses exclusively on this proprietary model, utilizing Gemini’s free quota. For the open-source models, a sliding-window approach with mean pooling was applied to address token input limits, 2048 for *EmbeddingGemma-300m* and 512 for *CodeBERT* and *GraphCodeBERT*. All models produced 768-dimensional vector representations for each instance, derived from the concatenation of the code diff and its corresponding commit message.

The classification task was treated as a binary problem, and the dataset was split into an 80% training set and a 20% test set. We evaluated multiple classifiers, including RandomForest (RF), Support Vector Classifier (SVC), DecisionTree (DT), and Multi-Layer Perceptron (MLP). These models were selected to represent a diverse range of learning paradigms: tree-based ensembles (RF, DT), margin-based optimization (SVC), and neural networks (MLP), which have been widely applied in various software engineering tasks [20, 64].

Table 4: Performance comparison of ML models trained on *text-embedding-004* embeddings. Metrics are calculated by targeting the buggy samples (*pos_label* = “Buggy”).

Classifier	Accuracy	Precision	Recall	F1-score	MCC
RF	0.816	0.841	0.82	0.83	0.629
SVC	0.884	0.918	0.866	0.891	0.768
DT	0.686	0.717	0.706	0.712	0.366
MLP	0.895	0.929	0.876	0.902	0.791

Table 4 presents the performance comparison of all models trained on *text-embedding-004* embeddings, showing that the *MLP Classifier* achieved the best results, with an accuracy of 0.895, F1-score of 0.902, and MCC of 0.791, outperforming other classifiers

in all metrics. Similar experiments were conducted using open-source embeddings; however, due to space limitations, only the MLP results are reported in Table 5, as it consistently outperformed other classifiers. Among the open-source models, *GraphCodeBERT* combined with MLP achieved the highest performance (F1-score = 0.831), exhibiting a modest gap of around 7% compared to the proprietary model. This suggests that while proprietary embeddings retain an advantage, open-source alternatives offer competitive and cost-effective performance for this task.

Table 5: Performance comparison of MLP classifier trained on different LLM embeddings. The first row of *text-embedding-004* is taken from Table 4 for comparison. Metrics are calculated by targeting the buggy samples (*pos_label* = “Buggy”).

Model	Accuracy	Precision	Recall	F1-score	MCC
text-embedding-004	0.895	0.929	0.876	0.902	0.791
EmbeddingGemma	0.80	0.815	0.82	0.818	0.593
CodeBERT	0.765	0.738	0.833	0.782	0.533
GraphCodeBERT	0.828	0.831	0.83	0.831	0.656

Furthermore, we evaluated the *MLPClassifier* using the *text-embedding-004* embeddings under a Leave-One-Out (LOO) strategy to test its robustness, where a single example is used as the test sample and the remaining serve as the training data. This approach not only offers more robust evaluation, but also can show if a model’s accuracy can be improved with more training samples. Encouragingly, the model achieved improved performance with an accuracy of 0.9036, precision of 0.8980, recall of 0.9141, F1-score of 0.906, and MCC of 0.8073, setting a new benchmark compared to prior best

results in RQ1 and RQ2. In terms of class-wise performance the model maintained a balanced detection capability, with precision and recall exceeding 0.89 for both classes and achieving F1-scores of 0.91 for *Buggy* and 0.9 for *NotBuggy*.

Summary of RQ3: Embedding-based ML models, particularly MLPClassifier, demonstrate strong predictive power for tangled change detection, achieving over 90% F1-score and outperforming prior baselines of RQ1 and RQ2. More encouragingly, our results suggest that the performance of embedding-based models can be improved with an enlarged training dataset.

4.4 RQ4: Potential impact of LLM-based untangling on future method-level bug prediction models

From the results of RQ1 to RQ3, we observe that the LLM-based method-level tangled change detection approach performs well with high accuracy and F1-scores using both prompting techniques and embedding-based ML models. This suggests that such approaches can effectively reduce noise in bug datasets. However, an important question remains: *Does this noise reduction benefit machine learning-based bug prediction models?* Building such a model, however, requires addressing several challenges. First, a broad set of code metrics is needed, whereas the dataset from Chowdhury et al. [20] includes only five. Second, Mashhadi et al. [54] demonstrate that combining code metrics with embeddings leads to better performance, requiring us to generate embeddings for all methods. Finally, a thorough evaluation across multiple algorithms and fine-tuning strategies is essential. Due to these requirements, constructing a full bug prediction model is beyond the scope of this paper. Nonetheless, if we can demonstrate that our noise reduction approach improves the ability of code metrics to distinguish between bug-prone and non-bug-prone methods, likely, the performance of machine learning models relying on these metrics will also improve. Therefore, this research question examines whether the distribution of code metrics becomes more distinguishable in the dataset produced by our method.

In RQ1 and RQ2, we evaluated multiple prompting strategies and LLMs, showing that *GPT-4o* combined with few-shot + chain-of-thought prompting delivers the best and balanced performance comparing all metrics. In RQ3, we examined the feasibility of using LLM-generated embeddings with downstream machine learning classifiers to detect tangled changes. Although the latter approach yields better results, it requires generating embeddings for all tangled code diffs from bug-fix commits to support prediction with the machine learning model, which is more complex and time-consuming than the few-shot + chain-of-thought approach. Therefore, we employed *GPT-4o* with few-shot + chain-of-thought prompting to detect tangled changes across all methods in 49 software projects in a fast and efficient manner.

In the dataset by Chowdhury et al. [20], each method is associated with its change history and the number of other methods modified in the same commit. We apply the few-shot and chain-of-thought strategy using *GPT-4o* to separate *NotBuggy* changes within bug-fix commits, thereby removing noise. To enable replication, the algorithm for creating the Less-Noisy dataset is included in the

shared repository (algorithm/algorithm.png). After constructing the *Less-Noisy* dataset, we obtain four types of methods:

Noisy NotBuggy. Following Chowdhury et al. [20], methods are labeled as *NotBuggy* if they are at least two years old and never involved in a bug-fix. This set is called *Noisy* because it comes from the original noisy dataset, which excludes non-buggy methods that were mixed with buggy changes.

Less-Noisy NotBuggy. It expands on the previous set by including methods confirmed as *NotBuggy* after LLM-based untangling.

Noisy Buggy. Following Chowdhury et al. [20], methods are labeled as *Buggy* if they appear in bug-fix commits, without accounting for tangling, hence the dataset is considered Noisy.

Less-Noisy Buggy. This set includes methods identified as truly buggy after applying our LLM-based untangling approach.

The dataset provided by Chowdhury et al. [20] contains five code metrics computed for each version of a method across its history. For our analysis, we consider the metrics from the first version of each method to ensure consistency across the dataset. These code metrics are defined as follows: *Size* refers to the number of source lines of code, excluding comments and blank lines [20, 21]; *Readability* is a score that reflects the ease of reading the source code, developed by Buse et al. [15]; *McCabe* measures cyclomatic complexity by counting the number of independent execution paths in a method [55]; *FanOut* represents the number of distinct methods called by a given method, indicating its dependency footprint [20, 64]; and *Maintainability Index (MI)* is a composite metric that combines several aspects, including complexity and size, to offer a comprehensive assessment of code maintainability [61].

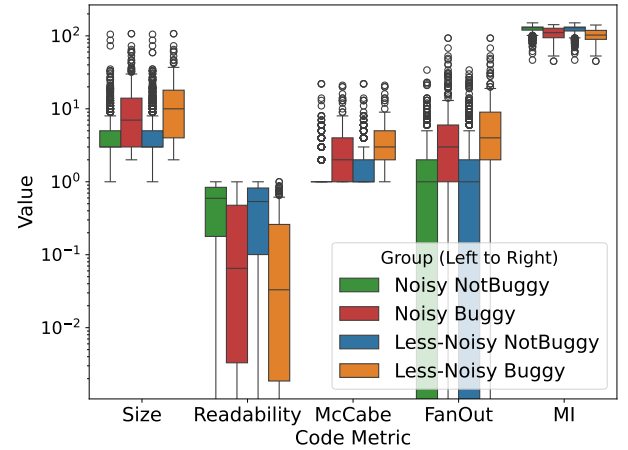


Figure 4: Distribution of code metrics for *Buggy* and *NotBuggy* methods in the *Titan* project across the *Noisy* and *Less-Noisy* datasets. Metric distributions differ more significantly in the *Less-Noisy* dataset.

Figure 4 presents the distribution of these 5 code metrics across the four method sets for the project *Titan*. Visually, the difference in *Size* between *Less-Noisy Buggy* and *Less-Noisy NotBuggy* methods appears larger than the other two sets. A similar pattern is observed for all other code metrics. To determine whether these differences are statistically significant, we use the Wilcoxon rank-sum test [56]. This non-parametric test is widely adopted in software engineering research due to its robustness in comparing distributions [6, 17, 19].

The results show that the differences in distributions are indeed statistically significant ($p \leq 0.05$) for all code metrics.

As the differences are statistically significant for both datasets, we calculated Cliff’s delta effect size [53] to understand the magnitude of these differences. This non-parametric measure is also widely used in software engineering research [6, 17, 19] and does not require assumptions about the data distribution. The results show that, in the *Noisy* dataset, the effect size is large for *MI* and medium for all other code metrics. In the *Less-Noisey* dataset, the effect size is large across all code metrics.

To evaluate the generalizability of our results, we performed Wilcoxon rank-sum tests across all 49 projects. In the *Noisy* dataset, significant differences ($p \leq 0.05$) were observed in 47 projects (95.92%) for *Size*, *McCabe*, and *MI*, 45 projects (91.84%) for *Readability*, and 48 projects (97.96%) for *FanOut*. In the *Less-Noisey* dataset, we observe improved results, with one additional project exhibiting statistically significant separation for each metric.

To further quantify the differences between *Buggy* and *NotBuggy* methods, we compute Cliff’s Delta effect sizes across all 49 projects. Table 6 presents the effect size distributions for both the *Noisy* and *Less-Noisey* datasets. We categorize the effect sizes following Hess et al. [38]: *Negligible* (< 0.147), *Small* ($0.147 \leq \delta < 0.33$), *Medium* ($0.33 \leq \delta < 0.474$), and *Large* ($\delta \geq 0.474$).

Table 6: Cliff’s Delta effect size values compare *Noisy* and *Less-Noisey* datasets, grouped by effect size category and shown as percentages in the format *Noisy / Less-Noisey*. The percentages for the *Noisy* dataset do not sum to 100% because the *RxJava* project was excluded as it had no *NotBuggy* methods.

Metric	Negligible	Small	Medium	Large
Size	0 / 0	14.29 / 6.12	46.94 / 20.41	36.73 / 73.47
Readability	2.04 / 2.04	36.73 / 24.49	42.86 / 40.82	16.33 / 32.65
McCabe	2.04 / 2.04	34.69 / 8.16	36.73 / 42.86	24.49 / 46.94
FanOut	0 / 0	10.20 / 6.12	46.94 / 22.45	40.82 / 71.43
MI	0 / 0	12.24 / 6.12	34.69 / 16.33	51.02 / 77.55

The table reports the proportion of projects falling into each effect size category for all five code metrics. The results indicate that, for most metrics, the *Less-Noisey* dataset shows a substantial shift toward larger effect sizes. For example, for the code metric, *Size*, in the *Less-Noisey* dataset, 73% of the projects exhibit a large effect size in the difference between *Buggy* and *NotBuggy* methods, compared to only 36% in the *Noisy* dataset. Similar patterns are observed for the other four code metrics. These findings clearly demonstrate that our filtering approach effectively eliminates noise and increases the distinction between the code metrics of buggy and non-buggy methods. As a result of this enhanced separability, we believe that our approach will support improved performance in future machine learning models for bug prediction.

Summary of RQ4: After constructing the *Less-Noisey* dataset using the LLM-based approach, statistical analysis reveals significantly improved separability (with larger effect sizes) in code metrics between buggy and non-buggy methods. This suggests that the noise reduction could enhance the performance of future ML-based bug prediction models.

5 Threats to Validity

Construct Validity. The manual annotation of code diffs as *Buggy* or *NotBuggy* may be affected by subjective bias. Although annotators were provided with explicit guidelines and a UI tool to assist in their evaluations, subjective judgment and semantic ambiguity in commit messages may have influenced labeling outcomes.

External Validity. The dataset used in this study comprises only open-source Java projects, which may limit the generalizability of our findings. Applying this approach to other programming languages, project domains, or commit conventions may require modifications to the prompts or retraining of models.

Internal Validity. The dataset used in this study was curated from method histories of Java projects obtained through CodeShovel [31], which is not fully accurate. Any inaccuracies in the method histories could impact the results.

Conclusion Validity. It is affected by all of the above mentioned threats.

6 Conclusion and Future Work

Method-level bug prediction is considered one of the holy grails in software engineering research. However, it remains an open challenge, primarily due to the absence of a noise-free method-level bug dataset [20, 64]. To address this, we evaluated the effectiveness of various LLMs and prompting strategies for detecting tangled changes at the method level. Our results show that even zero-shot LLMs achieve high accuracy when both commit messages and code diffs are provided (RQ1). Performance improves further when combining chain-of-thought with few-shot prompting (RQ2). Additionally, embedding-based models deliver even greater accuracy gains (RQ3). Across RQ1 to RQ3, open-source models also demonstrate strong performance. While they do not yet match proprietary models, they offer efficient alternatives with acceptable trade-offs. Building on these insights, we generated a *Less-Noisey* dataset using our LLM-based method. This dataset shows promise in developing more accurate method-level bug prediction models as it exhibits a significantly stronger power of commonly used code metrics to differentiate between bug-prone and non-bug-prone methods compared to the original *Noisy* dataset (RQ4).

The implications of our findings extend beyond the untangling problem and contribute to broader discussions on the capabilities and limitations of LLMs in software engineering. Although challenges such as hallucinations, reproducibility, and data leakage remain [41, 67, 84, 85], our study offers empirical evidence supporting the effectiveness of LLMs in tasks such as code reasoning, classification, and dataset construction [40, 60, 90]. In particular, our work shows that when guided by carefully designed prompts, LLMs can significantly improve software engineering workflows.

Looking ahead, a promising direction for future research is to enrich the dataset with additional code metrics and embeddings and to explore a range of machine learning models for building more effective method-level bug prediction models. To support continued progress, we have publicly released our code and dataset, allowing researchers to use them directly or adapt them to their specific needs.

References

- [1] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925* (2025).
- [2] Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Unified Pre-training for Program Understanding and Generation. *arXiv:2103.06333*.
- [3] Hadeel Alsolai, Marc Roper, and Dua Nassar. 2018. Predicting Software Maintainability in Object-Oriented Systems Using Ensemble Techniques. In *2018 IEEE International Conference on Software Maintenance and Evolution*. 716–721.
- [4] Xavier Amatriain. 2024. Prompt Design and Engineering: Introduction and Advanced Methods. *arXiv:2401.14423*.
- [5] Jaehyeon Bae, Seoryeong Kwon, and Seunghwan Myeong. 2024. Enhancing software code vulnerability detection using gpt-4o and claude-3.5 sonnet: A study on prompt engineering techniques. *Electronics* 13, 13 (2024), 2657.
- [6] Abdul Ali Bangash, Hareem Sahar, Abram Hindle, and Karim Ali. 2020. On the time-based conclusion stability of cross-project defect prediction models. *Empirical Software Engineering* 25, 6 (2020), 5047–5083.
- [7] V.R. Basili, L.C. Briand, and W.L. Melo. 1996. A Validation of Object-Oriented Design Metrics as Quality Indicators. *IEEE Transactions on Software Engineering* 22, 10 (1996), 751–761.
- [8] Samuel Silvestre Batista, Bruno Branco, Otávio Castro, and Guilherme Avelino. 2024. Code on Demand: A Comparative Analysis of the Efficiency Understandability and Self-Correction Capability of Copilot ChatGPT and Gemini. In *Proceedings of the XXIII Brazilian Symposium on Software Quality*. 351–361.
- [9] Robert M Bell, Thomas J Ostrand, and Elaine J Weyuker. 2011. Does measuring code change improve fault prediction?. In *Proceedings of the 7th international conference on predictive models in software engineering*. 1–8.
- [10] Ziqian Bi, Keyu Chen, Chiung-Yi Tseng, Danyang Zhang, Tianyang Wang, Hongying Luo, Lu Chen, Junming Huang, Jibin Guan, Junfeng Hao, et al. 2025. Is GPT-OSS Good? A Comprehensive Evaluation of OpenAI's Latest Open Source Models. *arXiv preprint arXiv:2508.12461* (2025).
- [11] Christian Bird, Adrian Bachmann, Eirik Aune, John Duffy, Abraham Bernstein, Vladimir Filkov, and Premkumar Devanbu. 2009. Fair and Balanced? Bias in Bug-Fix Datasets. In *Proceedings of the 7th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on The Foundations of Software Engineering*. 121–130.
- [12] Jürgen Börstler and Barbara Paech. 2016. The Role of Method Chains and Comments in Software Readability and Comprehension—An Experiment. *IEEE Transactions on Software Engineering* 42, 9 (2016), 886–898.
- [13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. 1877–1901.
- [14] Marc Bruni, Fabio Gabrielli, Mohammad Ghafari, and Martin Kropp. 2025. Benchmarking Prompt Engineering Techniques for Secure Code Generation with GPT Models. *arXiv:2502.06039*.
- [15] Raymond PL Buse and Westley R Weimer. 2009. Learning a metric for code readability. *IEEE Transactions on software engineering* 36, 4 (2009), 546–558.
- [16] Siyu Chen, Shengbin Xu, Yuan Yao, and Feng Xu. 2022. Untangling Composite Commits by Attributed Graph Clustering. In *Proceedings of the 13th Asia-Pacific Symposium on Internetware*. 117–126.
- [17] Yaohui Chen, Peng Li, Jun Xu, Shengjian Guo, Rundong Zhou, Yulong Zhang, Tao Wei, and Long Lu. 2020. Savior: Towards bug-driven hybrid testing. In *2020 IEEE Symposium on Security and Privacy (SP)*. 1580–1596.
- [18] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. *arXiv:2306.14924*.
- [19] Shaiful Chowdhury, Stephanie Borle, Stephen Romansky, and Abram Hindle. 2019. Greenscaler: training software energy models with automatic test generation. *Empirical Software Engineering* 24, 4 (2019), 1649–1692.
- [20] Shaiful Chowdhury, Gias Uddin, Hadi Hemmati, and Reid Holmes. 2024. Method-Level Bug Prediction: Problems and Promises. *ACM Transactions on Software Engineering and Methodology* 33, 4 (2024), 1–31.
- [21] Shaiful Alam Chowdhury, Gias Uddin, and Reid Holmes. 2022. An empirical study on maintainable method size in java. In *Proceedings of the 19th International Conference on Mining Software Repositories*. 252–264.
- [22] christine fisher. 2020. Boeing found another software bug on the 737 Max. <https://www.engadget.com/2020-02-06-boeing-737-max-software-bug.html> [Online; last accessed 2025-07-18].
- [23] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
- [24] Martin Dias, Alberto Bacchelli, Georgios Gousios, Damien Cassou, and Stéphane Ducasse. 2015. Untangling Fine-Grained Code Changes. In *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering*. 341–350.
- [25] Mengdan Fan, Wei Zhang, Haiyan Zhao, Guangtai Liang, and Zhi Jin. 2024. Detect Hidden Dependency to Untangle Commits. In *2024 39th IEEE/ACM International Conference on Automated Software Engineering*. 179–190.
- [26] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, et al. 2020. Codebert: A pre-trained model for programming and natural languages. *arXiv preprint arXiv:2002.08155* (2020).
- [27] Emanuel Giger, Marco D'Ambros, Martin Pinzger, and Harald C. Gall. 2012. Method-Level Bug Prediction. In *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. 171–180.
- [28] Yossi Gil and Gal Lalouche. 2017. On the Correlation between Size and Metric Validity. *Empirical Software Engineering* 22 (2017), 1–27.
- [29] Louie Giray. 2023. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering* 51, 12 (2023), 2629–2633.
- [30] Google DeepMind. 2024. Advancing Gemini: December 2024 Update. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/> [Online; last accessed: 2025-07-11].
- [31] Felix Grund, Shaiful Alam Chowdhury, Nick C. Bradley, Braxton Hall, and Reid Holmes. 2021. CodeShovel: Constructing Method-Level Source Code Histories. In *2021 IEEE/ACM 43rd International Conference on Software Engineering*. 1510–1522.
- [32] Bo Guo and Myoungkyu Song. 2017. Interactively Decomposing Composite Changes to Support Code Review and Regression Testing. In *2017 IEEE 41st Annual Computer Software and Applications Conference*, Vol. 1. 118–127.
- [33] Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, et al. 2020. Graphcodebert: Pre-training code representations with data flow. *arXiv preprint arXiv:2009.08366* (2020).
- [34] Hideaki Hata, Osamu Mizuno, and Tohru Kikuno. 2012. Bug Prediction Based on Fine-Grained Module Histories. 200–210.
- [35] Steffen Herbold, Alexander Trautsch, Benjamin Ledel, Alireza Aghamohammadi, Taher A. Ghaleb, Kuljit Kaur Chahal, Tim Bossenmaier, Bhavet Nagaria, Philip Makedonski, Matin Nili Ahmabadi, Kristof Szabados, Helge Spieker, Matej Madeja, Nathaniel Hoy, Valentina Lenarduzzi, Shangwen Wang, Gema Rodriguez-Pérez, Ricardo Colomo-Palacios, Roberto Verdecchia, Paramvir Singh, Yihao Qin, Debasish Chakraborti, Willard Davis, Vijay Walunj, Hongjun Wu, Diego Marclio, Omar Alam, Abdullah Aldaei, Idan Amit, Burak Turhan, Simon Eismann, Anna-Katharina Wickert, Ivano Malavolta, Matus Sulir, Fatemeh Fard, Austin Z. Henley, Stratos Kourtzanidis, Eray Tuzun, Christoph Treude, Simin Maleki Shamasbi, Ivan Pashchenko, Marvin Wyrich, James Davis, Alexander Serebrenik, Ella Albrecht, Ethem Utku Aktas, Daniel Strüder, and Johannes Erbel. 2022. A Fine-Grained Data Set and Analysis of Tangling in Bug Fixing Commits. *Empirical Software Engineering* 27, 6 (2022), 125.
- [36] Kim Herzig, Sascha Just, and Andreas Zeller. 2016. The Impact of Tangled Code Changes on Defect Prediction Models. *Empirical Software Engineering* 21, 2 (2016), 303–336.
- [37] Kim Herzig and Andreas Zeller. 2013. The Impact of Tangled Code Changes. In *2013 10th Working Conference on Mining Software Repositories*. 121–130.
- [38] Melinda R Hess and Jeffrey D Kromrey. 2004. Robust confidence intervals for effect sizes: A comparative study of Cohen's d and Cliff's delta under non-normality and heterogeneous variances. In *annual meeting of the American Educational Research Association*, Vol. 1.
- [39] Tiancheng Hu and Nigel Collier. 2024. Quantifying the Persona Effect in LLM Simulations. *arXiv:2402.10811*.
- [40] Imen Jaoua, Oussama Ben Sghaier, and Houari Sahraoui. 2025. Combining Large Language Models with Static Analyzers for Code Review Generation. In *2025 IEEE/ACM 22nd International Conference on Mining Software Repositories*. 174–186.
- [41] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (2023), 248:1–248:38.
- [42] Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. Beyond Words: A Comparative Analysis of LLM Embeddings for Effective Clustering. In *Advances in Intelligent Data Analysis XXII*. 205–216.
- [43] Hiroyuki Kirinuki, Yoshiki Higo, Keisuke Hotta, and Shinji Kusumoto. 2014. Hey! Are You Committing Tangled Changes?. In *Proceedings of the 22nd International Conference on Program Comprehension*. 262–265.
- [44] Pavneet Singh Kochhar, Yuan Tian, and David Lo. 2014. Potential Biases in Bug Localization: Do They Matter?. In *Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering*. 803–814.
- [45] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
- [46] Jinhyuk Lee, Zhu Yun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. 2024. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327* (2024).

- [47] Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same Task, More Tokens: The Impact of Input Length on the Reasoning Performance of Large Language Models. arXiv:2402.14848.
- [48] Yi Li, Shaohua Wang, and Tien N. Nguyen. 2022. UTANGO: Untangling Commits with Context-Aware, Graph-Based, Code Change Clustering Learning Model. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 221–232.
- [49] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9 (2023), 195:1–195:35.
- [50] Xiaoyu Liu, LiGuo Huang, Chuanyi Li, and Vincent Ng. 2018. Linking Source Code to Untangled Change Intents. In *2018 IEEE International Conference on Software Maintenance and Evolution*. 393–403.
- [51] Yu Liu, Duantengchuan Li, Kaili Wang, Zhuoran Xiong, Fobo Shi, Jian Wang, Bing Li, and Bo Hang. 2024. Are LLMs Good at Structured Outputs? A Benchmark for Evaluating Structured Output Capabilities in LLMs. *Information Processing & Management* 61, 5 (2024), 103809.
- [52] Wei Ma, Shangqing Liu, Mengjie Zhao, Xiaofei Xie, Wenhong Wang, Qiang Hu, Jie Zhang, and Yang Liu. 2024. Unveiling code pre-trained models: Investigating syntax and semantics capacities. *ACM Transactions on Software Engineering and Methodology* 33, 7 (2024), 1–29.
- [53] Guillermo Macbeth, Eugenia Razumiejczyk, and Rubén Daniel Ledesma. 2011. Cliff’s Delta Calculator: A non-parametric effect size program for two groups of observations. *Universitas Psychologica* 10, 2 (2011), 545–555.
- [54] Ehsan Mashhadi, Hossein Ahmadvand, and Hadi Hemmati. 2023. Method-level bug severity prediction using source code metrics and LLMs. In *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*. 635–646.
- [55] Thomas J McCabe. 1976. A complexity measure. *IEEE Transactions on software Engineering* 4 (1976), 308–320.
- [56] Patrick E McKnight and Julius Najab. 2010. Mann-Whitney U Test. *The Corsini encyclopedia of psychology* (2010), 1–1.
- [57] T. Menzies, J. Greenwald, and A. Frank. 2007. Data Mining Static Code Attributes to Learn Defect Predictors. *IEEE Transactions on Software Engineering* 33, 1 (2007), 2–13.
- [58] Ran Mo, Shaozhi Wei, Qiong Feng, and Zengyang Li. 2022. An Exploratory Study of Bug Prediction at the Method Level. *Information and Software Technology* 144 (2022), 106794.
- [59] Ward Muylaert and Coen De Roover. 2018. [Research Paper] Untangling Composite Commits Using Program Slicing. In *2018 IEEE 18th International Working Conference on Source Code Analysis and Manipulation*. 193–202.
- [60] Daye Nam, Andrew Macvean, Vincent Helleboom, Bogdan Vasilescu, and Brad Myers. 2024. Using an LLM to Help With Code Understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–13.
- [61] Paul Oman and Jack Hagemeister. 1992. Metrics for assessing a software system’s maintainability. In *Proceedings Conference on Software Maintenance 1992*. 337–338.
- [62] OpenAI. 2025. GPT-4o mini: Advancing Cost-Efficient Intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/> [Online; last accessed: 2025-07-11].
- [63] Profir-Petru Părtăchi, Santanu Kumar Dash, Miltiadis Allamanis, and Earl T. Barr. 2020. Flexeme: Untangling Commits Using Lexical Flows. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 63–74.
- [64] Luca Pascarella, Fabio Palomba, and Alberto Bacchelli. 2020. On the Performance of Method-Level Bug Prediction: A Negative Result. *Journal of Systems and Software* 161 (2020), 110493.
- [65] Alina Petukhova, João P. Matos-Carvalho, and Nuno Fachada. 2024. Text Clustering with Large Language Model Embeddings.
- [66] Md Saidur Rahman and Chanchal K. Roy. 2017. On the Relationships Between Stability and Bug-Proneness of Code Clones: An Empirical Study. In *2017 IEEE 17th International Working Conference on Source Code Analysis and Manipulation*. 131–140.
- [67] June Sallou, Thomas Durieux, and Annibale Panichella. 2024. Breaking the Silence: The Threats of Using LLMs in Software Engineering. In *Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*. 102–106.
- [68] Francisco Servant and James A. Jones. 2017. Fuzzy Fine-Grained Code-History Analysis. In *2017 IEEE/ACM 39th International Conference on Software Engineering*. 746–757.
- [69] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On Second Thought, Let’s Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. arXiv:2212.08061.
- [70] Bo Shen, Wei Zhang, Christian Kästner, Haiyan Zhao, Zhao Wei, Guangtai Liang, and Zhi Jin. 2021. SmartCommit: A Graph-Based Interactive Assistant for Activity-Oriented Commits. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 379–390.
- [71] Thomas Shippey, Tracy Hall, Steve Counsell, and David Bowes. 2016. So You Need More Method Level Datasets for Your Software Defect Prediction? Voilà! (*ESEM ’16*).
- [72] Mifta Sintaha, Noor Nashid, and Ali Mesbah. 2023. Katana: Dual slicing based context for learning bug fixes. *ACM Transactions on Software Engineering and Methodology* 32, 4 (2023), 1–27.
- [73] sixsintex. 2024. Most expensive software bugs in history: Sixsintex. <https://www.sixsintex.com/insights/ten-most-expensive-bugs-in-history-part-1> [Online; last accessed 2025-07-18].
- [74] Shiyu Sun, Yanhui Li, Lin Chen, Yuming Zhou, and Jianhua Zhao. 2025. Boosting Code-line-level Defect Prediction with Spectrum Information and Causality Analysis. In *2025 IEEE/ACM 47th International Conference on Software Engineering*. 776–776.
- [75] Yida Tao and Sunghun Kim. 2015. Partitioning Composite Code Changes to Facilitate Code Review. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*. 180–190.
- [76] Michele Tufano, Cody Watson, Gabriele Bavota, Massimiliano Di Penta, Martin White, and Denys Poshyvanyk. 2019. An empirical study on learning bug-fixing patches in the wild via neural machine translation. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 28, 4 (2019), 1–29.
- [77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30.
- [78] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. 2025. Embeddinggemma: Powerful and lightweight text representations. arXiv preprint arXiv:2509.20354 (2025).
- [79] Zixu Wang, Weiyuan Tong, Peng Li, Guixin Ye, Hao Chen, Xiaoqing Gong, and Zhanyong Tang. 2023. BugPre: An Intelligent Software Version-to-Version Bug Prediction System Using Graph Convolutional Neural Networks. *Complex & Intelligent Systems* 9, 4 (2023), 3835–3855.
- [80] Supatsara Wattanakriengkrai, Patanamon Thongtanunam, Chakkrit Tantithamthavorn, Hideaki Hata, and Kenichi Matsumoto. 2022. Predicting Defective Lines Using a Model-Agnostic Technique. *IEEE Transactions on Software Engineering* 48, 5 (2022), 1480–1496.
- [81] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652.
- [82] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.
- [83] Sheng-Bin Xu, Si-Yu Chen, Yuan Yao, and Feng Xu. 2025. Detecting and Untangling Composite Commits via Attributed Graph Modeling. *Journal of Computer Science and Technology* 40, 1 (2025), 119–137.
- [84] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2024. LLM Lies: Hallucinations Are Not Bugs, but Features as Adversarial Examples. arXiv:2310.01469.
- [85] Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive Mirage: A Review of Hallucinations in Large Language Models. arXiv:2309.06794.
- [86] Shouyu Yin, Shikai Guo, Hui Li, Chenchen Li, Rong Chen, Xiaochen Li, and He Jiang. 2025. Line-Level Defect Prediction by Capturing Code Contexts With Graph Convolutional Networks. *IEEE Transactions on Software Engineering* 51, 1 (2025), 172–191.
- [87] Qunhong Zeng, Yuxia Zhang, Zhiqing Qiu, and Hui Liu. 2025. A First Look at Conventional Commits Classification. In *Proceedings of the IEEE/ACM 47th International Conference on Software Engineering*. 2277–2289.
- [88] Qunhong Zeng, Yuxia Zhang, Zeyu Sun, Yujie Guo, and Hui Liu. 2024. COLARE: Commit Classification via Fine-grained Context-aware Representation of Code Changes. In *2024 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 752–763.
- [89] Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. 2024. ∞ Bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15262–15277.
- [90] Tianming Zheng, Haojun Liu, Hang Xu, Xiang Chen, Ping Yi, and Yue Wu. 2024. Few-VulD: A Few-shot learning framework for software vulnerability detection. *Computers & Security* 144 (2024), 103992.
- [91] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625 (2022).
- [92] Thomas Zimmermann, Rahul Premraj, and Andreas Zeller. 2007. Predicting Defects for Eclipse. In *Third International Workshop on Predictor Models in Software Engineering*. 9–9.