

# Agentic Refactoring: An Empirical Study of AI Coding Agents

**KOSEI HORIKAWA**, Nara Institute of Science and Technology, Japan

**HAO LI**, Queen's University, Canada

**YUTARO KASHIWA**, Nara Institute of Science and Technology, Japan

**BRAM ADAMS**, Queen's University, Canada

**HAJIMU IIDA**, Nara Institute of Science and Technology, Japan

**AHMED E. HASSAN**, Queen's University, Canada

Agentic coding tools, such as OpenAI Codex, Claude Code, and Cursor, are transforming the software engineering landscape. These AI-powered systems function as autonomous teammates capable of planning and executing complex development tasks. Agents have become active participants in refactoring, a cornerstone of sustainable software development aimed at improving internal code quality without altering observable behavior. Despite their increasing adoption, there is a critical lack of empirical understanding regarding how agentic refactoring is utilized in practice, how it compares to human-driven refactoring, and what impact it has on code quality.

To address this empirical gap, we present a large-scale study of AI agent-generated refactorings in real-world open-source Java projects, analyzing 15,451 refactoring instances across 12,256 pull requests and 14,998 commits derived from the AIDev dataset. Our empirical analysis shows that refactoring is a common and intentional activity in this development paradigm, with agents explicitly targeting refactoring in 26.1% of commits. Analysis of refactoring types reveals that agentic efforts are dominated by low-level, consistency-oriented edits, such as `CHANGE VARIABLE TYPE` (11.8%), `RENAME PARAMETER` (10.4%), and `RENAME VARIABLE` (8.5%), reflecting a preference for localized improvements over the high-level design changes common in human refactoring. Additionally, the motivations behind agentic refactoring focus overwhelmingly on internal quality concerns, with maintainability (52.5%) and readability (28.1%). Furthermore, quantitative evaluation of code quality metrics shows that agentic refactoring yields small but statistically significant improvements in structural metrics, particularly for medium-level changes, reducing class size and complexity (e.g., Class LOC median  $\Delta = -15.25$ ).

**CCS Concepts:** • **Software and its engineering → Maintaining software; Automatic programming; Software evolution.**

**Additional Key Words and Phrases:** Agentic Coding, Coding Agent, Refactoring, Pull Requests, Large Language Models

## ACM Reference Format:

Kosei Horikawa, Hao Li, Yutaro Kashiwa, Bram Adams, Hajimu Iida, and Ahmed E. Hassan. 2025. Agentic Refactoring: An Empirical Study of AI Coding Agents. 1, 1 (November 2025), 23 pages. <https://doi.org/XXXX-XXXXXX>

---

Authors' Contact Information: **Kosei Horikawa**, horikawa.kosei.hk1@naist.ac.jp, Nara Institute of Science and Technology, Ikoma, Japan; **Hao Li**, Queen's University, Kingston, Canada, hao.li@queensu.ca; **Yutaro Kashiwa**, Nara Institute of Science and Technology, Ikoma, Japan, yutaro.kashiwa@is.naist.jp; **Bram Adams**, Queen's University, Kingston, Canada, bram.adams@queensu.ca; **Hajimu Iida**, Nara Institute of Science and Technology, Ikoma, Japan, iida@itc.naist.jp; **Ahmed E. Hassan**, Queen's University, Kingston, Canada, ahmed@cs.queensu.ca.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM XXXX-XXXX/2025/11-ART

<https://doi.org/XXXXXX.XXXXXXX>

## 1 Introduction

Refactoring is a cornerstone of sustainable software development that improves a software system's internal quality without changing its observable behavior [36]. Since Martin Fowler's catalog [19] established the field's conceptual foundation, subsequent studies have emphasized refactoring as a core maintenance activity that supports long-term software evolvability. Prior work links refactoring to improved readability [26, 45], higher developer productivity [33], and the removal of code smells [52], as well as to preparing a codebase for future modifications [39].

Despite its importance, refactoring demands deep domain knowledge and specialized skills. When executed improperly, refactoring can introduce bugs [38], break existing tests [23, 40], or destabilize the system [6]. To mitigate these inherent risks, researchers have proposed many sophisticated approaches over the decades. These techniques include methods to identify appropriate refactoring locations [56], development of refactoring recommendation systems [32], and automatic refactoring approaches [48]. While several studies reported that practitioners did not use refactoring tools [34], the advent of large language models (LLMs) has recently accelerated these techniques and spurred practical adoption of automatic refactoring [5, 17, 50]. Many studies report that practitioners now actively leverage LLMs to generate and apply refactoring operations [3, 16, 58, 60].

The software engineering landscape is now undergoing another fundamental transformation with the rise of agentic coding tools [21]. Unlike traditional prompt-based LLM workflows, where developers manually guide the AI step-by-step, agentic coding tools such as OpenAI Codex,<sup>1</sup> Claude Code,<sup>2</sup> and Cursor<sup>3</sup> operate as AI teammates [28]. These agents can autonomously plan, execute, test, and iterate on complex development tasks with minimal human intervention [44]. In this new paradigm, agents act as collaborative refactoring teammates, actively participating in the refactoring process, ranging from simple cleanups to substantial design modifications. However, while these tools are being rapidly adopted, there is currently no empirical understanding of how this new class of agentic refactoring is used in practice, how it compares to human-driven refactoring, and what impact it has on code quality.

To address these critical gaps, we conduct a large-scale empirical study of refactorings generated by AI agents in real-world open-source projects based on the AIDev dataset [28]. We identify 15,451 refactoring instances across 12,256 pull requests and 14,998 commits in Java generated by agentic coding tools. We examine these agentic refactorings along four core dimensions (i.e., prevalence, types, purposes, and impacts) to answer the following research questions (RQs):

**RQ<sub>1</sub>: To what extent is refactoring activity in agentic commits?** To determine whether refactoring is a significant and intentional activity in agentic software development, we first quantify how often it occurs. We find that refactoring is common, appearing in 26.1% of agentic Java commits, which contain a total of 7,127 refactoring instances.

**RQ<sub>2</sub>: What are the common types of agentic refactoring?** To assess the current sophistication of AI agents, it is crucial to determine whether they perform simple cosmetic cleanups or engage in more complex structural transformations. Our analysis reveals that agentic refactoring is dominated by low-level, consistency-oriented edits like renaming and type changes, showing a clear preference for localized improvements over the high-level design changes common in human refactoring.

**RQ<sub>3</sub>: What is the purpose of agentic refactoring?** Understanding the why behind these operations is essential to see if developers use agents for proactive quality improvement or for

<sup>1</sup><https://chatgpt.com/features/codex>

<sup>2</sup><https://www.claude.com/product/claude-code>

<sup>3</sup><https://cursor.com/>

other tactical reasons. We find that agentic refactoring is overwhelmingly driven by the desire to improve internal code quality, specifically maintainability (52.5%) and readability (28.1%).

**RQ4: To what extent does agentic refactoring affect code quality?** To justify agentic refactoring adoption, it is essential to quantitatively measure whether agentic refactoring provides tangible benefits to the codebase's structural health. The results show that agentic refactoring yields small but significant improvements in structural metrics like code size and complexity, yet it currently fails to reduce the overall count of known design and implementation smells.

Our findings provide the first large-scale empirical baseline of agentic refactoring. These results have direct implications for developers, helping them set expectations for what current agents can (and cannot) reliably refactor. For researchers, our work opens new avenues for studying human-agent collaboration in software maintenance and provides a foundation for future studies. Finally, for coding agent builders, our findings highlight current limitations (e.g., focus on low-level consistency, limited impact on design smells) and opportunities for creating agents that can perform more sophisticated, structurally-aware refactorings.

**Replication Package.** To facilitate replication and further studies, we provide the data used in our replication package.<sup>4</sup>

**Paper Structure.** The remainder of this paper is organized as follows. Section 2 presents the motivating examples. Section 3 details our study design, data collection, and analysis methodology. Section 4 reports the empirical findings. Section 5 summarizes the key insights and discusses the implications of our study. Section 6 reviews related work. Section 7 outlines threats to validity, and Section 8 concludes the paper.

## 2 Motivating Example

The integration of AI agents into software development workflows has accelerated rapidly in recent years. Coding agents such as Claude Code, Devin, and Cursor can not only generate code but also collaboratively refactor existing codebases with developers [61]. Rather than acting as fully autonomous systems, these agents increasingly function as collaborative refactoring teammates, assisting humans in restructuring large or complex code to enhance maintainability and readability.

Figure 1a shows a representative example of an agent refactoring a long, complex method into multiple smaller, reusable helper methods.<sup>5</sup> The agent automatically introduces helper methods such as `printUsageAndExit()` and `execute()` to reduce method length and clarify control flow. At first glance, this transformation clearly enhances readability and maintainability by reducing cognitive complexity and adhering to modular design principles. However, it also raises a deeper question: Do agentic refactorings like these consistently improve software quality metrics (e.g., lines of code, cyclomatic complexity, coupling), or do they simply restructure code superficially?

In contrast, Figure 1b shows another type of agentic refactoring, where an agent systematically renames variables to improve consistency.<sup>6</sup> Such operations improve naming clarity and stylistic uniformity but have little direct effect on structural quality metrics. This contrast highlights a key uncertainty in modern agentic software development: while agents can execute both low-level syntactic refactorings (e.g., renaming) and high-level structural ones (e.g., method extraction), their true impact on measurable code quality remains unclear.

Together, these examples illustrate that AI agents are becoming increasingly active participants in the refactoring process, performing tasks ranging from simple cleanups to substantial design modifications. Yet, despite their growing presence, there remains limited empirical understanding of

<sup>4</sup>[https://github.com/Mont9165/Agent\\_Refactoring\\_Analysis](https://github.com/Mont9165/Agent_Refactoring_Analysis)

<sup>5</sup><https://github.com/jeffreyjose07/power-of-gman/pull/3/files>

<sup>6</sup><https://github.com/Ddemon26/2006Scape/pull/4/files>

**Refactor long methods into smaller ones #3**

Merged jeffreyjose07 merged 1 commit into main from codex/refactor-long-methods-into-smaller-ones on Jul 1

Conversation 0 Commits 1 Checks 0 Files changed 2

```
src/main/java/com/example/geektrust/Main.java +49 -41
@@ -41,16 +41,23 @@ public Main(PathFindingStrategy pathFinder) {
41
42     public static void main(String[] args) {
43         if (args.length == 0) {
44             System.err.println("Error: No input file provided.");
45             System.err.println("Usage: java -jar <jar-file> <input-file>");
46             System.exit(ERROR_NO_INPUT);
47     }
48
49     Main app = new Main();
50     try {
51         app.processInputFile(args[0]);
52     } catch (Exception e) {
53         e.printStackTrace();
54     }
55 }
```

(a) Decomposing a long method into helper methods to improve readability and reduce complexity.

**[BOT] refactor(rename): Game audio methods #4**

Merged Ddemon26 merged 13 commits into master from codex/rename-game.java-class-chunk on Jun 29

Conversation 0 Commits 13 Checks 0 Files changed 6 +228 -226

Changes from all commits File filter Conversations

Filter changed files

```
2006Scape Client/src/main/java/Class56.java
@@ -1,18 +1,18 @@
1   abstract class Class56 {
2
3   -   abstract void method827(int i, byte[] is, int i_2_, boolean bool);
4
5   -   abstract void method828();
6
7   -   public Class56() {
8   -       /* empty */
9
10  +   abstract void playMidi(int i, byte[] is, int i_2_, boolean bool);
11
12  +   abstract void shutdown();
```

(b) Standardizing variable names across multiple files.

Fig. 1. Examples of agentic refactoring.

how such agentic refactorings are used in practice. Specifically, it is not yet known how frequently AI agents participate in refactoring activities, what types of refactoring operations they most commonly perform, why these refactorings are initiated, or how they affect internal software quality metrics.

To answer these questions, we conduct a large-scale empirical study of agentic refactorings in real-world open-source projects. We systematically examine four dimensions of agentic refactoring: the **frequency** of AI participation (RQ1), the **types** of refactoring operations performed (RQ2), the **purpose** expressed by developers and AI agents (RQ3), and the resulting **impacts** on internal quality metrics (RQ4). Together, these research questions aim to clarify whether agentic refactorings merely automate surface-level cleanup tasks or genuinely contribute to structural and maintainability improvements in software systems.

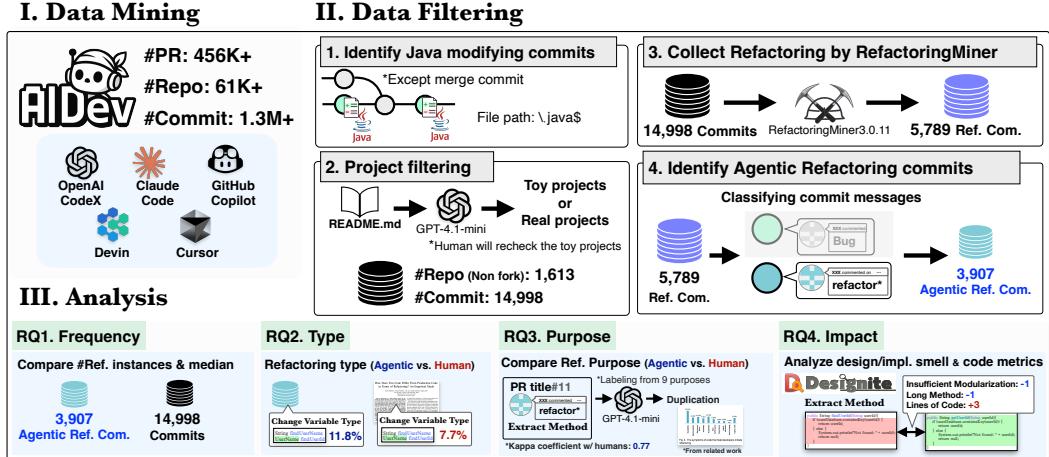


Fig. 2. Overview of the study design.

### 3 Data Collection

In this section, we describe the data mining and data filtering as outlined in Figure 2.

#### 3.1 Data Mining

To study agentic refactoring in real-world projects, we start with the AIDev dataset [28]. This dataset is an ideal source as it contains 932,791 pull requests (PRs) created by five coding agents from over 61,000 repositories. However, the AIDev dataset only includes commit data for repositories with more than 100 GitHub stars. To gather a more comprehensive set of data points, we leverage the GitHub REST API<sup>7</sup> to mine all commits based on the provided metadata. In total, we collect 1,311,057 agentic commits.

#### 3.2 Data Filtering

To construct a focused dataset tailored for refactoring analysis, we perform a multi-stage filtering process based on the collected data curated from the AIDev dataset.

**3.2.1 Identify Java Modifying Commits.** In this study, we focus on commits that modify at least one .java file to leverage mature and well-supported research tools such as RefactoringMiner [1, 57] and DesigniteJava [49]. In addition, to avoid duplicated changes and ambiguous parent relationships, we exclude all merge commits.

**3.2.2 Project Filtering.** To ensure that our analysis targets substantive software systems rather than educational or trivial repositories, we apply several steps to filter the projects.

**Step 1 – Automated Classification.** We leverage GPT-4.1-mini to classify each repository based on the content of its README.md file. Each project is assigned to one of the following four categories:

- **production\_grade**: Actively maintained or widely used software intended for real-world use (e.g., applications, libraries, developer tools).
- **specialized\_project**: Niche, experimental, or research-oriented repositories that still provide meaningful software functionality.

<sup>7</sup><https://docs.github.com/en/rest>

- `toy_or_example`: Tutorial repositories, coursework material, evaluation harnesses, or other minimal or demonstrative examples.
- `uncertain`: Insufficient or unclear project description that prevents reliable classification (e.g., minimal or missing `README.md`).

*Step 2 – Manual Verification.* Since automated classification may incorrectly identify some meaningful projects as toy examples, we manually reviewed all repositories initially labeled as `toy_or_example`. Among these, we identified 7 misclassified projects, which are reclassified as `production_grade` (5 projects) and `specialized_project` (2 projects), respectively.

*Step 3 – Final Project Selection.* After classification and verification, we retain 1,134 projects labeled as `production_grade` and 501 projects labeled as `specialized_project`. We excluded 1,235 repositories categorized as `toy_or_example` and 362 as `uncertain`. Finally, to avoid redundancy, we remove forked repositories. This results in a final corpus of 1,613 unique software projects for analysis. The Java subset includes 14,998 non-merge commits suitable for automated refactoring detection and metric extraction. This filtering step reduces noise from trivial repositories and ensures that our analysis focuses on meaningful and substantive software systems.

**3.2.3 Collect Refactoring Operations by RefactoringMiner.** To identify specific refactoring operations within our set of Java commits, we use RefactoringMiner 3.0<sup>8</sup> for each commit to detect specific refactoring operations. RefactoringMiner detects 103 distinct refactoring types and identifies their precise locations, achieving an overall F-score of 99.5% [1, 57]. Specifically, RefactoringMiner analyzes the abstract syntax trees (ASTs) of the modified Java files between consecutive revisions to identify the types and locations of applied refactorings.

This process generates a dataset linking specific commits to the refactoring operations they contain. From the initial set of 14,998 commits, RefactoringMiner identifies 5,789 commits as containing at least one refactoring operation.

**3.2.4 Identify Agentic Refactoring Commits.** To understand how often agentic refactoring is an *intentional* act, we have to identify commits that explicitly state refactoring intent. We label a commit as agentic refactoring if (i) RefactoringMiner detects at least one refactoring, and (ii) the commit message signals refactoring intent. To identify this intent, we use common keywords and patterns (e.g., “refactor”, “cleanup”, “restructure”) directly adapted from prior work on self-affirmed refactoring [2]. The complete list of patterns is presented in Table 1. Applying this procedure to the 14,998 refactoring commits yields 3,907 agentic refactoring commits.

**3.2.5 Dataset Overview.** After filtering, our curated dataset contains 14,998 unique, non-merge commits that modify at least one Java file. Of these, 3,907 are labeled as `agentic_refactoring_commits` (those with explicit refactoring intent in their messages) and the remaining 11,091 commits categorized as `other_commits`.

These 14,998 commits originate from 12,256 PRs across 1,613 repositories. Among these PRs, 11,504 (93.9%) are closed and 10,645 (86.9%) are merged. This high merge rate indicates that most agentic contributions were integrated into their respective projects, demonstrating substantial acceptance of agent-generated code. Regarding agent participation, the distribution across the five agents is shown in Table 2. The data reveal that OpenAI Codex dominates the dataset, accounting for 89.3% of all commits and 94.3% of PRs. Devin and Cursor contribute 5.7% and 2.8% respectively, while Claude Code accounts for only 0.6% of the commits analyzed.

---

<sup>8</sup><https://github.com/tsantalis/RefactoringMiner/releases/tag/3.0.11>

Table 1. List of Self-Affirmed Refactoring (SAR) Pattern from AlOmar *et al.* [2]

(1) Refactor*	(30) Removed poor coding practice	(59) Change design
(2) Mov*	(31) Improve naming consistency	(60) Modularize the code
(3) Split*	(32) Removing unused classes	(61) Code cosmetics
(4) Fix*	(33) Pull some code up	(62) Moved more code out of
(5) Introduce*	(34) Use better name	(63) Remove dependency
(6) Decompos*	(35) Replace it with	(64) Enhanced code beauty
(7) Reorganiz*	(36) Make maintenance easier	(65) Simplify internal design
(8) Extract*	(37) Code cleanup	(66) Change package structure
(9) Merg*	(38) Minor Simplification	(67) Use a safer method
(10) Renam*	(39) Reorganize project structures	(68) Code improvements
(11) Chang*	(40) Code maintenance for refactoring	(69) Minor enhancement
(12) Restructur*	(41) Remove redundant code	(70) Get rid of unused code
(13) Reformat*	(42) Moved and gave clearer names to	(71) Fixing naming convention
(14) Extend*	(43) Refactor bad designed code	(72) Fix module structure
(15) Remov*	(44) Getting code out of	(73) Code optimization
(16) Replac*	(45) Deleting a lot of old stuff	(74) Fix a design flaw
(17) Rewrit*	(46) Code revision	(75) Nonfunctional code cleanup
(18) Simplifi*	(47) Fix technical debt	(76) Improve code quality
(19) Creat*	(48) Fix quality issue	(77) Fix code smell
(20) Improv*	(49) Antipattern bad for performances	(78) Use less code
(21) Add*	(50) Major/Minor structural changes	(79) Avoid future confusion
(22) Modif*	(51) Clean up unnecessary code	(80) More easily extended
(23) Enhanc*	(52) Code reformatting & reordering	(81) Polishing code
(24) Rework*	(53) Nicer code / formatted / structure	(82) Move unused file away
(25) Inlin*	(54) Simplify code redundancies	(83) Many cosmetic changes
(26) Redesign*	(55) Added more checks for quality factors	(84) Inlined unnecessary classes
(27) Cleanup	(56) Naming improvements	(85) Code cleansing
(28) Reduc*	(57) Renamed for consistency	(86) Fix quality flaws
(29) Encapsulat*	(58) Refactoring towards nicer name analysis	(87) Simplify the code

Table 2. Distribution of commits and PRs by associated AI agent.

AI Agent	# PRs (%)	# Commits (%)
OpenAI Codex	11,557 (94.3%)	13,389 (89.3%)
Devin	335 (2.7%)	860 (5.7%)
Cursor	337 (2.8%)	663 (4.4%)
Claude Code	27 (0.2%)	86 (0.6%)
Total	12,256 (100%)	14,998 (100%)

## 4 Results

In this section, we provide the motivation, approach, and findings for each of our research questions. Figure 2 provides an overview of our study design.

### 4.1 RQ<sub>1</sub>: To what extent is refactoring activity in agentic commits?

4.1.1 *Motivation.* Agentic coding is transitioning from a novelty to an everyday practice, yet little is known about the extent of refactoring performed by these agents in real-world projects. Although agentic coding tools promise to act as autonomous collaborators rather than mere code generators [21], most research has focused on their ability to create new features or fix bugs. Their role in software maintenance, particularly refactoring, remains largely unexplored. Before assessing the quality or types of refactorings agents perform, we must first establish a baseline: Is agentic refactoring a significant real-world activity or merely an incidental byproduct of other development tasks? Quantifying its frequency is the first step toward understanding whether these agents genuinely contribute to refactoring.

Table 3. Summary of detected refactoring instances in agentic commits.

	# Commits (%)	# Instances (%)
Agentic Refactoring	3,907 (26.1%)	7,127 (46.1%)
Others	11,091 (73.9%)	8,324 (53.9%)
Total	14,998 (100%)	15,451(100%)

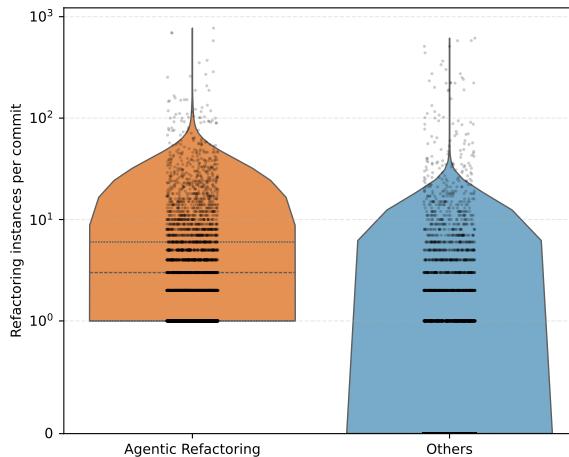


Fig. 3. Distribution of refactoring instances per refactoring commit (agentic vs. others).

**4.1.2 Approach.** We record the distribution of refactoring instances for the collected 14,998 commits. To determine whether the distributions of refactoring instances in agentic refactorings and other commits are significantly different, we perform the Mann-Whitney U test [30] at a significance level of  $\alpha = 0.05$ . We also compute Cliff's delta  $d$  [29] effect size to quantify the difference based on the following thresholds [42]:

$$\text{Effect size} = \begin{cases} \text{negligible}, & \text{if } |d| \leq 0.147 \\ \text{small}, & \text{if } 0.147 < |d| \leq 0.33 \\ \text{medium}, & \text{if } 0.33 < |d| \leq 0.474 \\ \text{large}, & \text{if } 0.474 < |d| \leq 1 \end{cases} \quad (1)$$

**4.1.3 Findings.** **Finding #1: Refactoring is common in agentic software development and often appears even without explicit intent.** As shown in Table 3, agentic refactoring commits account for 26.1% (3,907 out of 14,998) of all commits and contain 7,127 detected refactoring instances. We also find 8,324 refactoring instances in other commits, whose messages do not explicitly indicate refactoring. This pattern suggests tangled commits: while implementing a feature or fixing a bug, agents also refactor nearby code (e.g., rename variables, extract small helpers) within the same commit.

**Finding #2: When agents state refactoring intent, they perform significantly more refactoring instances than other commits.** Figure 3 shows the distribution of refactoring instances per commit in agentic refactoring and other commits. The result of the Mann-Whitney U test shows that the difference is statistically significant ( $p \leq 0.001$ ). In addition, the result of Cliff's

delta ( $d = 0.838$ ) indicates a large effect size. The results suggest that when AI agents explicitly signal refactoring intent, they perform more concentrated and substantial restructuring activities. In contrast, when refactoring is unacknowledged, it tends to appear sporadically or incidentally as part of broader code modifications, such as feature implementation or bug fixing.

### Answer to RQ1

Refactoring is a common and intentional activity in agentic software development, with 26.1% of agentic commits explicitly targeting refactoring.

## 4.2 RQ<sub>2</sub>: What are the common types of agentic refactoring?

**4.2.1 Motivation.** Refactoring encompasses a wide range of activities, from simple, localized cleanups such as variable renaming to complex, structural transformations like splitting a class or extracting a superclass, which require a comprehensive understanding of the software system. Analyzing the types of refactorings performed by agents serves as a direct indicator of their sophistication and capability. Are these agents merely acting as “code janitors,” automating low-level syntactic cleanup, or are they beginning to function as “software architects,” executing deep structural design improvements essential for long-term maintainability? By comparing the distribution of agentic refactoring types with known human refactoring patterns [22], this RQ establishes the first empirical profile of agentic refactoring and evaluates its current level of maturity.

**4.2.2 Approach.** For each detected refactoring instance, RefactoringMiner provides its specific type from a set of 103 distinct operations. We record these types for all Agentic Refactoring instances in our dataset. To understand the structural impact of these operations, we classify them using the three abstraction levels proposed by Murphy-Hill *et al.* [34], based on the structural impact of the refactoring instances. Since Murphy-Hill *et al.* [34] did not classify all 103 refactoring types detected by RefactoringMiner, we extended their classification framework by applying the same criteria to categorize the remaining refactoring types. Specifically, we classified each refactoring based on whether it modifies (i) only signatures or interfaces, (ii) both signatures and code blocks, or (iii) only internal code blocks. The details of the three abstraction levels of refactoring instances are as below:

- **High-level (58 types):** Refactorings that only change the signatures of classes, methods, or fields without modifying internal code blocks. These alter the public interface and often require changes to calling code (e.g., RENAME METHOD, ADD PARAMETER, MOVE CLASS, CHANGE METHOD ACCESS MODIFIER).
- **Medium-level (21 types):** Refactorings that change both signatures and code blocks, bridging internal logic and external structure (e.g., EXTRACT METHOD, INLINE METHOD, MOVE AND INLINE METHOD, CHANGE ATTRIBUTE TYPE).
- **Low-level (24 types):** Refactorings confined exclusively to code blocks (typically within method bodies) that are not visible externally (e.g., RENAME VARIABLE, CHANGE VARIABLE TYPE, EXTRACT VARIABLE, REPLACE ANONYMOUS WITH LAMBDA).

The complete mapping of all 103 refactoring types to abstraction levels is available in our replication package.<sup>9</sup> We then calculate the distribution of refactoring instances across these three levels. To identify differences between agent and human behavior, we compare our results with human refactoring patterns reported in prior work [22], mapping the human-driven instances to the same three-level framework for a direct comparison.

<sup>9</sup>[https://github.com/Mont9165/Agent\\_Refactoring\\_Analysis/blob/main/Refactoring\\_Level\\_Classification.md](https://github.com/Mont9165/Agent_Refactoring_Analysis/blob/main/Refactoring_Level_Classification.md)

Table 4. Refactoring abstraction levels: Agentic Refactoring vs Human

Abstraction Level	% Agentic Refactoring	% Human Refactoring [22]
High-level	43.0%	54.9%
Medium-level	21.2%	20.7%
Low-level	35.8%	24.4%

Table 5. Comparison of Agentic and Human Refactoring Instances (Top 3 at each level)

Abstraction Level	Actor	Refactoring Type	% Instances
High-level	Agent	<b>RENAME ATTRIBUTE</b>	6.0%
		ADD METHOD ANNOTATION	4.1%
		<b>CHANGE METHOD ACCESS MODIFIER</b>	3.8%
	Human	ADD METHOD ANNOTATION	5.8%
		<b>RENAME METHOD</b>	4.4%
		<b>ADD PARAMETER</b>	4.3%
Medium-level	Agent	<b>MOVE AND INLINE METHOD</b>	7.2%
		CHANGE ATTRIBUTE TYPE	5.2%
		CHANGE PARAMETER TYPE	2.5%
	Human	CHANGE PARAMETER TYPE	6.2%
		<b>CHANGE RETURN TYPE</b>	4.7%
		CHANGE ATTRIBUTE TYPE	3.5%
Low-level	Agent	CHANGE VARIABLE TYPE	11.8%
		RENAME PARAMETER	10.4%
		RENAME VARIABLE	8.5%
	Human	CHANGE VARIABLE TYPE	7.7%
		RENAME VARIABLE	3.3%
		RENAME PARAMETER	3.0%

4.2.3 *Findings. Finding #3: Agentic refactoring emphasizes low-level edits more than humans, while performing fewer high-level structural changes.* As shown in Table 4, agents perform fewer high-level refactorings (signatures) compared to humans (43.0% versus 54.9%). This difference is offset by a substantial increase in low-level refactorings (code blocks), which account for 35.8% of agent refactorings compared to 24.4% for humans. In contrast, the proportion of medium-level refactorings (both signatures and code blocks) is nearly identical for agents (21.2%) and humans (20.7%). These results suggest that current AI agents tend to perform refactoring in a more localized and conservative manner, focusing on internal method bodies rather than broader system interfaces.

**Finding #4: Agents and humans share a common focus on low-level refactoring operations but diverge in high-level refactoring.** Table 5 highlights distinct agentic priorities in the most common refactoring operations. At the low level, agents and humans are aligned, as the top three operations are identical for both: CHANGE VARIABLE TYPE, RENAME PARAMETER, and RENAME VARIABLE. Agents perform these naming and type consistency refactorings even more frequently, with these

three types alone accounting for 30.7% of all agentic refactoring instances. At the high level, however, their focus diverges. While ADD METHOD ANNOTATION is a shared high-frequency operation, humans frequently perform major API changes such as RENAME METHOD (4.4%) and ADD PARAMETER (4.3%). In contrast, agents prioritize RENAME ATTRIBUTE (6.0%) and CHANGE METHOD ACCESS MODIFIER (3.8%), indicating a preference for modifying class-level state and visibility rather than method-level APIs. At the medium level, both agents and humans frequently perform CHANGE ATTRIBUTE TYPE and CHANGE PARAMETER TYPE, but their top operations differ: agents most often apply MOVE AND INLINE METHOD (7.2%), while humans prioritize CHANGE RETURN TYPE (4.7%).

### Answer to RQ2

Agentic refactoring is dominated by lower-level refactorings (35.8%) such as type changes and renaming, occurring more often than in human refactoring (24.4%). In contrast, agents perform fewer high-level refactorings (43.0%) than humans (54.9%), while medium-level refactorings occur at similar rates. Overall, current agents primarily focus on localized improvements rather than architectural changes.

## 4.3 RQ<sub>3</sub>: What is the purpose of agentic refactoring?

**4.3.1 Motivation.** In human-driven development, intent plays a central role. Developers refactor code to improve readability, reduce coupling in preparation for new features, or make the code easier to test [26]. But what about agents? This research question explores the motivations behind agentic refactoring. Do developers use agents as proactive “quality partners” to enhance maintainability and reduce technical debt, or as tactical assistants for short-term objectives such as improving the readability of a specific method? Understanding these motivations is essential for aligning agent behavior with developer expectations.

**4.3.2 Approach.** To analyze why agentic refactoring is performed, we classify the primary motivation of each Agentic Refactoring commit into one of ten motivation categories derived from prior work by Kim *et al.* [26] (e.g., maintainability, readability, testability, logical mismatch *etc.*). Although Kim *et al.* originally include *logical mismatch* as a category, we exclude it in this study because it corresponds to bug fixing rather than refactoring motivation. For consistency with this modified scheme, we also rerank the human refactoring data reported by Kim *et al.* [26] according to our adjusted category definitions. The classification uses the Pull Request title, commit message, and detected refactoring types as input. Due to the large sample size of the collected Agentic Refactoring commits, we leverage GPT-4.1-mini to automatically categorize each commit.

To validate the reliability of the automatic classification, two inspectors (each with seven years of programming experience) independently label a stratified sample of Agentic Refactoring commits (ten per category). We assess inter-rater reliability using Cohen’s  $\kappa$  coefficient [18], which measures the degree of agreement between annotators. As shown in Table 6, inter-rater agreement between the two human annotators is excellent (Cohen’s  $\kappa = 0.83$ ). Disagreements are resolved through discussion with another inspector (seventeen years of programming experience), producing the final reference labels. We then compare the GPT-4.1-mini annotations with these adjudicated human labels. The agreement remains excellent (Cohen’s  $\kappa = 0.77$ ), indicating that GPT-4.1-mini provides reliable motivation classification for large-scale empirical analysis.

**4.3.3 Findings. Finding #5: Agentic refactoring is primarily driven by maintainability, a purpose that is far more dominant for agents than for humans.** As shown in Figure 4, maintainability is the primary driver for more than half (52.5%) of all agentic refactoring commits.

Table 6. Reliability of refactoring purpose classification

Comparison	Cohen's $\kappa$	Accuracy	Macro F1
Human vs Human	0.83	0.85	0.85
GPT-4.1-mini vs Human	0.77	0.80	0.80

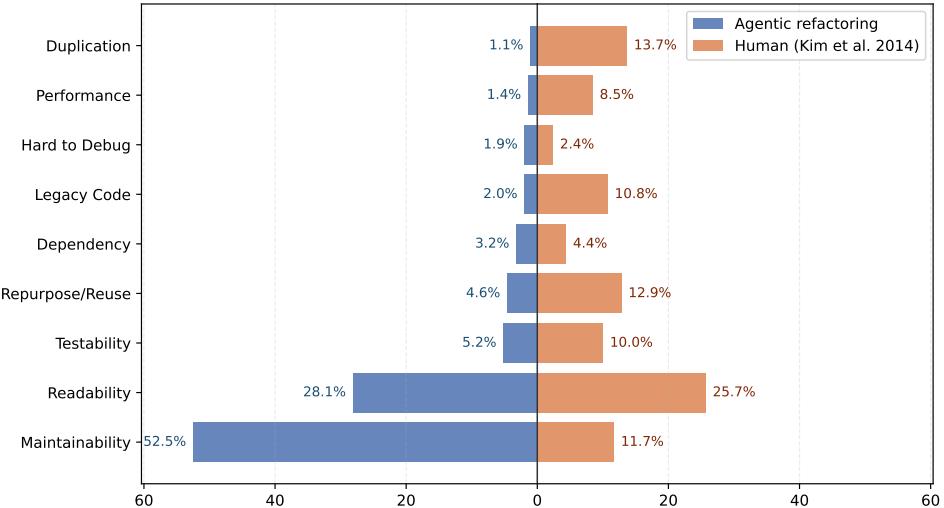


Fig. 4. Refactoring purpose comparison between agents and humans (normalized) [26]

This presents a stark contrast to human refactoring patterns, where maintainability is a much less frequent purpose (11.7%) [26]. The second most common agent purpose is readability (28.1%), which aligns closely with the most common human purpose (25.7%), establishing it as a shared, core priority for both. Together, these two categories account for over 80% of all agentic refactoring purposes. This suggests that developers primarily dispatch agents to perform day-to-day codebase care (e.g., cleanups and consistency edits) rather than to pursue broader design goals.

**Finding #6: Agentic refactoring rarely targets design-level improvements, such as duplication or dependency management.** In sharp contrast to the focus on maintainability, motivations related to design-level refactoring are negligible in agentic commits. For example, duplication (1.1%) and repurpose/reuse (4.6%) are among the least common agent motivations. This is the inverse of human-driven refactoring, where these two categories are prominent (13.7% and 12.9%, respectively), indicating that humans frequently refactor to improve modularity and reduce redundancy. This pattern suggests that current agents focus on localized quality cleanup rather than complex, system-wide restructuring.

**Finding #7: Corrective refactoring occasionally appears but remains a secondary purpose.** A small fraction of agentic refactoring is motivated by corrective concerns. This includes addressing code that is hard to debug (1.9%) or dealing with legacy code (2.0%). These instances often involve tactical improvements rather than deep bug fixing, such as adding logging (e.g., `log.debug()`)

where it was missing,<sup>10</sup> or modernizing code by upgrading Java versions (e.g., from Java 7 to Java 17).<sup>11</sup> These cases are present but clearly not the primary use case for agents.

### Answer to RQ3

Agentic refactoring is overwhelmingly motivated by internal quality concerns. Maintainability (52.5%) and readability (28.1%) are the dominant drivers, accounting for over 80% of all instances. This focus on localized cleanup contrasts with human refactoring, which more frequently addresses design-level concerns like duplication and code reuse, motivations that are rare in agentic commits.

## 4.4 RQ4: To what extent does agentic refactoring affect code quality?

**4.4.1 Motivation.** The core premise of refactoring is to improve a software system’s internal code quality. However, decades of research on human refactoring indicate that this outcome is not guaranteed. Refactorings may fail to remove existing smells or even introduce new ones [6, 8], and similar uncertainty applies to agents. To justify delegating maintenance tasks to AI agents, we need quantitative evidence of their impact. This RQ shifts the focus from intent to outcomes. By analyzing established software metrics and comparing smell counts before and after agentic commits, we provide the first empirical assessment of whether agentic refactoring improves code quality.

**4.4.2 Approach.** To quantify the impact of refactoring on internal code quality, we extract object-oriented metrics and design smells before and after each Agentic Refactoring commit using DesigniteJava [49]. We select DesigniteJava because it is widely used in empirical software engineering research (e.g., [27, 47]). As it performs static analysis directly on source code, it ensures consistent metric extraction across heterogeneous projects without requiring build configuration. The tool computes class- and method-level metrics commonly used in software quality assessment [10], including size (*Lines of Code*), complexity (*Cyclomatic Complexity*, *Weighted Methods per Class*), coupling (*Fan-in*, *Fan-out*), cohesion (*Lack of Cohesion in Methods*), and inheritance depth (*Depth of Inheritance Tree*). DesigniteJava also detects 27 design and implementation smells (e.g., *Long Method*, *Complex Method*, *Cyclic Dependency*), enabling a multi-perspective analysis of code maintainability.

In addition, we evaluate before-and-after changes ( $\Delta = \text{after} - \text{before}$ ) in internal code quality metrics detected by DesigniteJava for agentic refactoring commits. Before-and-after changes (i.e., values of  $\Delta$ ) are aggregated along two dimensions: (1) the abstraction level (high, medium, low) of the refactoring instances, and (2) the purpose category of the refactoring instances. To study the differences in these code-quality metrics changes, we perform the Wilcoxon signed-rank test [62], with  $p$ -values adjusted via the Benjamini–Hochberg FDR [20]. We also compute the rank-biserial effect size to quantify the difference. For cross-group comparisons (e.g., purposes or levels), we use Kruskal–Wallis tests [15] with FDR adjustment, and when relevant, we summarize post-hoc contrasts descriptively. Given the skewed distributions of deltas, we also report the median values of these before-and-after changes.

**4.4.3 Findings. Finding #8: Agentic refactoring yields statistically significant but not practically significant (i.e., negligible effect size) reductions in smell counts.** Figure 5 shows that the before-and-after distributions for both design and implementation smells almost completely

<sup>10</sup><https://github.com/Braindeiko01/CRduels/commit/4ea01106f0cf74663cf1dccecd94b4475e11794a>

<sup>11</sup><https://github.com/BrettMiller99/ReallyOldJavaProject/commit/daea563191184899891e1a1c66e2e8d60fa94bfc>

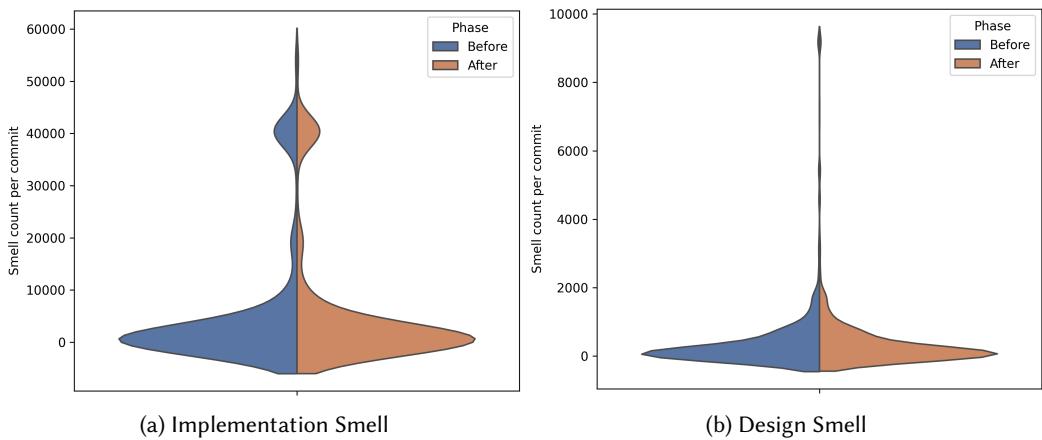


Fig. 5. Smell Count Distribution (Before vs. After)

overlap, indicating no visible shift in typical smell levels. Consistently, the median values remain virtually identical for both smell categories (Design: before 77.50, after 77.50; Implementation: before 355.50, after 356.00), resulting in a median  $\Delta$  of 0.00 in both cases. Although the Wilcoxon signed-rank test detects a statistically significant difference (FDR-adjusted  $p < 0.001$ ), the effect sizes are negligible (Cohen's  $d = -0.027$  and  $-0.026$ ). The mean  $\Delta$  values ( $-6.63$  and  $-32.85$  for Design and Implementation Smells, respectively) suggest that meaningful reductions occur only in a small subset of commits. In other words, AI agents can remove smells, but they do so inconsistently, and improvements are concentrated in exceptional cases rather than being routine.

**Finding #9: Agentic refactoring improves structural code metrics, whereas smell counts show no significant reduction.** As summarized in Table 7, several class-level metrics improve after agentic refactoring. Specifically, Class LOC (median  $\Delta = -15.25$ ), WMC (median  $\Delta = -2.07$ ), Fan-Out (median  $\Delta = -0.02$ ), and Fan-In (median  $\Delta = -0.01$ ) all show FDR-adjusted significance with negligible-to-small effect sizes. These results indicate that agentic refactoring tends to simplify structural complexity (size and coupling) more reliably than it eliminates higher-level smell patterns.

**Finding #10: Quality improvements are largest for medium-level refactorings, modest for low-level edits, and minimal for high-level signature-only changes.** Grouping by abstraction level, medium-level refactorings produce the most consistent gains (Table 7): Class LOC and WMC drop meaningfully (medians  $\Delta = -15.25$  and  $-2.07$ ), and method-level LOC decreases (median  $\Delta = -1.79$ ). Low-level edits also reduce method LOC (median  $\Delta = -0.42$ ) but show a slight increase in method cyclomatic complexity (median  $\Delta = +0.08$ ), suggesting that localized refactoring may sometimes split code without reducing decision complexity. High-level signature-only changes show near-zero medians across most metrics, which is consistent with interface adjustments rather than structural simplification. Kruskal–Wallis tests across levels (FDR-adjusted) support these differences at  $\alpha = 0.05$ .

**Finding #11: Structural refactorings (e.g., decomposition and modularization) produce the largest quality improvements.** Analyzing median metric deltas by refactoring type reveals that refactorings introducing new structure or decomposing responsibilities lead to the most substantial quality gains. For example, EXTRACT SUBCLASS yields large median reductions in Class LOC ( $\Delta = -87.5$ ) and WMC ( $\Delta = -11.5$ ), while SPLIT CLASS similarly reduces Class LOC ( $\Delta = -16.0$ )

Table 7. Median  $\Delta$  (after–before) of structural metrics across refactoring abstraction levels in agentic refactoring commits. Negative values indicate improvement.

Metric	Low-level (code block)	Medium-level (signature + block)	High-level (signature)
Class-Level – Depth of Inheritance Tree*	–	0.00	0.01
Class-Level – Fan-Out*	–	-0.02	0.00
Class-Level – Fan-In*	–	-0.01	0.00
Class-Level – Number of Methods*	–	-0.10	0.01
Class-Level – Weighted Methods per Class*	–	-2.07	0.03
Class-Level – Lines of Code*	–	-15.25	0.10
Method-Level – Parameter Count*	0.00	0.01	0.11
Method-Level – Cyclomatic Complexity*	0.08	0.01	0.00
Method-Level – Lines of Code*	-0.42	-1.79	0.11

– Not applicable: Class-Level metrics are not measured for Low-level (code block) refactorings.

and WMC ( $\Delta = -4.0$ ).<sup>12</sup> These refactorings improve modularity and reduce class-level complexity by distributing responsibilities across more cohesive components. In contrast, refactorings that preserve overall structure, such as signature adjustments or local rewrites, tend to produce smaller metric shifts. These findings indicate that the impact of agentic refactoring is strongly influenced by refactoring type. While structural decomposition refactorings drive measurable improvements, localized edits have a more limited effect on structural indicators.

**Finding #12: Not all refactoring types improve measured metrics; some primarily support clarity or evolution.** Several high-frequency agentic refactoring types (e.g., identifier renames, access or annotation adjustments) show negligible before-and-after change in the analyzed structural metrics after FDR correction. This does not imply that these refactorings lack value; rather, their main benefits (e.g., readability, naming consistency, API clarity) are not captured by the selected design-level indicators. By contrast, transformation types with both semantic and structural effects may yield mixed outcomes. For example, MOVE AND INLINE METHOD tends to reduce method LOC (median  $\Delta = -0.5$ ) but can leave cyclomatic complexity unchanged, or even increase it slightly when logic is redistributed across helper methods. These observations emphasize that metric-based evaluation should be interpreted together with developer intent (see Section 4.3). Agentic refactoring is often used for code consistency and comprehension, and measurable structural improvements emerge primarily when maintainability is the explicit objective.

#### Answer to RQ4

Agentic refactoring yields statistically significant but generally small structural improvements, most notably for medium-level changes that reduce class size and complexity (e.g., Class LOC median  $\Delta = -15.25$ , WMC median  $\Delta = -2.07$ ). Design and implementation smell counts do not show FDR-significant reductions. Purpose matters: maintainability-oriented refactorings (e.g., EXTRACT SUBCLASS, SPLIT CLASS) produce clearer metric gains, whereas readability-oriented refactorings (e.g., renames) rarely affect structural indicators.

<sup>12</sup>[https://github.com/Mont9165/Agent\\_Refactoring\\_Analysis/blob/main/outputs/designite/group\\_heatmaps/refactoring\\_heatmap\\_sar\\_median\\_delta.csv](https://github.com/Mont9165/Agent_Refactoring_Analysis/blob/main/outputs/designite/group_heatmaps/refactoring_heatmap_sar_median_delta.csv)

## 5 Implications

In this section, we discuss the implications of our findings for researchers, developers, and coding agent builders.

### 5.1 Implications for Researchers

**Researchers should investigate the hidden burden of implicit refactoring on development workflows.** Section 4.1 shows that a majority of refactoring instances (53.9%) occur in commits without explicit refactoring intent (*i.e.*, tangled commits). This suggests that agents frequently perform refactorings as a side effect of other tasks. Such tangling can increase review effort, as developers must validate both the primary task (*e.g.*, a feature) and the incidental refactorings to ensure behavior is preserved. Prior work indicates that developers want to verify logic preservation [46], so a high volume of implicit refactoring could lengthen reviews and reduce trust. Future work should quantify this burden by analyzing review comments, merge latency, and revert rates for commits with implicit refactoring.

**Researcher should create benchmarks for high-level refactoring.** Since agents underperform at high-level refactorings versus humans (Section 4.2), the community needs curated benchmarks and gold standards for architectural refactorings (*e.g.*, EXTRACT CLASS, INTRODUCE PARAMETER OBJECT, MOVE CLASS). Datasets should include ground-truth intent, behavior-preserving tests, and expected post-conditions (reduced WMC/fan-out).

**Researchers should study the disconnect between agentic refactoring and measured quality outcomes.** The majority (86.9%) of agentic PRs are merged into the software projects (Section 3.2.5), and our findings reveal that 26.1% of agentic commits contain explicitly stated refactoring (Section 4.1). However, traditional quality metrics show minimal improvement (Section 4.4). Researchers should investigate this gap to understand whether these agentic refactorings (*e.g.*, renaming variables) provide tangible benefits to human developers, such as improved comprehension or reduced cognitive load, or whether agents are engaging in low-value “code churn” that fails to address deeper structural issues.

**Researchers should validate long-term quality outcomes through longitudinal analysis for agentic refactorings.** Our before–after analysis of structural metrics (Section 4.4) shows small but statistically significant improvements, particularly for medium-level changes (*e.g.*, Class LOC median  $\Delta = -15.25$ , WMC median  $\Delta = -2.07$ ). The core premise of refactoring, however, concerns long-term evolvability and maintainability. Although agentic refactoring is overwhelmingly motivated by maintainability (Section 4.3), it does not consistently reduce design and implementation smells. Researchers should assess whether these structural gains translate into long-term benefits by conducting longitudinal studies that track defect density, post-release defects, and effort for future modifications.

### 5.2 Implications for Developers

**Developers should leverage coding agents strategically based on refactoring abstraction levels.** Our analysis in Section 4.2 shows that AI agents excel at low-level refactorings compared to humans (35.8% vs. 24.4%) but lag behind in high-level design changes (43.0% vs. 54.9%). Agentic refactoring is dominated by low-level edits, such as CHANGE VARIABLE TYPE (11.8%), RENAME PARAMETER (10.4%), and RENAME VARIABLE (8.5%). Developers can delegate routine cleanup to agents and focus human effort on design-level restructuring that requires domain knowledge and architectural intent. Given that many refactoring instances (53.9%) occur implicitly within non-refactoring commits (Section 4.1), developers must remain vigilant during code review to validate these tangled changes.

### 5.3 Implications for Coding Agent Builders

**Coding agent builders should reduce refactoring tangling commits with commit hygiene policies.** Over half (53.9%) of refactoring instances appear in commits without explicit refactoring intent (Section 4.1), which increases review burden. Agent builders should guide agents to separate refactorings from major tasks (e.g., features or fixes), auto-split PRs by change intent, and batch consistency edits into self-contained commits. Generated PR summaries should clearly state intent (e.g., maintainability or readability) to build reviewer trust.

**Coding agent builders should evolve agents from tactical cleanup partners to autonomous architectural planners.** As shown in Section 4.2, agentic refactorings are dominated by low-level operations (e.g., CHANGE VARIABLE TYPE and RENAME PARAMETER) and underrepresent high-level refactoring compared to humans (43.0% vs. 54.9%). Agent builders should train the agents (or the backend LLM) on curated datasets of successful high-level architectural refactorings. They should also incorporate behavior checks, including tests and differential builds, to validate refactorings.

**Coding agent builders should equip agents with specialized tools to autonomously detect and fix design flaws.** A key finding in Section 4.4 is that agentic refactoring fails to consistently reduce the overall count of design and implementation smells (median  $\Delta$  of 0.00). This suggests agents are overlooking specific flaws. A practical way to boost agentic refactoring is to equip agents with refactoring-specific analysis tools (e.g., DesigniteJava) and feed their findings back into the planning loop so the agent can actively seek out smells and verify improvements. Exposing these tools through a model context protocol (MCP) lets agents autonomously decide when to run a smell scan, retrieve metrics, and re-invoke transformations until targets are met (e.g., lower WMC or fan-out).

## 6 Related Work

In this section, we discuss related work about the foundations of refactoring, large-scale empirical analyses, automated refactoring techniques, and AI-assisted software development.

### 6.1 Refactoring and Its Empirical Foundations

Refactoring is a disciplined code transformation activity aimed at improving internal software quality without modifying observable behavior [36]. Since Fowler’s catalog of refactoring patterns [19] established its conceptual foundation, subsequent studies have emphasized refactoring as a core maintenance practice that supports long-term software evolvability [9, 14, 24, 33, 63]. Refactorings often aim to improve readability, reduce duplication, and facilitate maintainability of software systems [25, 52]. Empirical work has revealed that developers refactor not only to address code smells but also to improve comprehension, reduce cognitive load, and prepare code for future changes [14]. For instance, Kim *et al.* [26] surveyed Microsoft developers and found that refactoring is widely regarded as beneficial for improving code quality and productivity. Specifically, developers cited the benefits of refactoring as improved readability, improved maintainability, and fewer bugs. Their quantitative analysis of Windows 7 modules also showed that a dedicated refactoring effort led to measurable benefits, including a significant reduction in inter-module dependencies and post-release defects. Their findings provide the conceptual basis for our analysis of refactoring purpose in agentic commits (RQ3). Building on these foundational insights, subsequent research has investigated refactoring practices at scale to understand their prevalence and impact on software quality.

## 6.2 Large-Scale Refactoring Studies and Quality Impact

The large-scale mining of refactorings has revealed that refactoring is pervasive in both industrial and open-source development [34, 59, 64]. More recently, large-scale refactoring detectors such as REFACTORMINER [1, 57] and REFDIFF [51] have enabled mining-based studies at scale.

Although refactoring is generally expected to improve quality, empirical evidence on its actual impact remains mixed. Some studies report positive effects on code stability, readability, and productivity, particularly in agile settings [22, 25, 33], whereas others show that refactoring does not consistently remove design problems or prevent defects [7]. For example, Cedrim *et al.* [8] found that less than 10% of refactorings effectively remove code smells, while over 30% introduce new ones. Similarly, Szoke *et al.* [54] observed that only large-scale, systematic bulk refactorings lead to measurable quality improvement, suggesting that scope matters as much as type. Bavota *et al.* [6] further noted that inheritance-related refactorings are especially error-prone, emphasizing the need to consider technical context.

Beyond the technical dimension, developer intent plays a crucial role. Palomba *et al.* [37] showed that refactorings frequently co-occur with bug fixes, implying opportunistic rather than proactive behavior. This aligns with Silva *et al.* [52] and Pantiuchina *et al.* [39], who found that evolving requirements, readability, and maintainability—rather than smell removal—drive most refactoring decisions. Taken together, these findings highlight that refactoring impact depends on its type, scope, and motivation—a perspective we extend to agentic refactoring in RQ2 and RQ4.

## 6.3 Automated Refactoring Support

Automated and semi-automated refactoring has been explored in software engineering for over two decades. Rule-based tools such as JDEODORANT [31, 55] and search-based refactoring engines [35] attempted to reduce manual effort; however, they suffered from low adoption due to lack of trust and limited semantic reasoning [34, 59]. Recently, the refactoring capability of large language models (LLMs) has also been explored [12, 17]. Cordeiro *et al.* [12] evaluated the refactoring quality of StarCoder2 under various prompt-engineering strategies (e.g., zero-shot, one-shot) and found that LLMs can outperform developers in removing code smells under specific prompting configurations. However, these studies focus on *prompt-based refactoring*, where models perform one-shot transformations given explicit instructions.

In contrast, our study explores *agentic refactoring*, where autonomous coding agents (e.g., OpenAI Codex, Devin, Cursor) plan, execute, and validate changes through iterative reasoning and feedback. This paradigm differs from prompt-based generation in that agents can decompose complex objectives, perform refactoring alongside other maintenance activities, and generate verifiable pull requests with minimal human intervention. Thus, our work provides the first large-scale empirical view of intentional, agentic refactoring practices carried out through agentic coding workflows.

## 6.4 AI-Assisted Software Development

AI-based coding assistance has evolved rapidly from autocomplete-style tools [53] to modern LLM-based copilots capable of nontrivial synthesis and transformation [13]. Empirical studies report that AI-generated code accelerates development but often introduce maintainability, redundancy, and security concerns [4, 43]. Trust and validation remain key challenges when adopting AI-generated code. Sergeyuk *et al.* [46] reported that developers frequently verify AI suggestions manually due to concerns over correctness and control. In particular, 21.9% of respondents avoided using AI for refactoring tasks, highlighting that trust is critical where functional behavior must remain unchanged. LLMs further exhibit inconsistency and limited contextual understanding [4, 43]. Even with identical prompts, tools like ChatGPT may produce divergent or unnecessary edits, increasing

review effort and cognitive overhead [17]. Evidence from agentic coding platforms further supports these findings. A recent study of autonomously generated pull requests (Agentic-PRs) [61] found that 45.1% required post-review fixes—most often for bugs, refactoring, or documentation—despite agents performing refactoring more frequently than humans (24.9% vs. 14.9%). Overall, while AI-assisted tools can restructure code effectively, human oversight remains essential to ensure correctness and maintainability—an issue that becomes even more significant in *agentic refactoring*, where AI autonomously plans and applies code changes.

## 6.5 Agentic Software Development and Research Gap

Recent work has introduced the notion of *agentic software engineering* [21], where AI systems act as autonomous collaborators capable of proposing, modifying, and integrating pull requests [28]. The AIDev dataset enables empirical research on such activities at scale. However, while early work analyzes AI contributions in issue resolution [11] and documentation [41], no study has examined AI participation in refactoring, nor how AI-generated refactorings differ in frequency, type, intent, and impact compared to human refactoring. Our work addresses this gap by providing the first large-scale empirical analysis of refactoring in agentic commits, introducing *Agentic Refactoring* as a novel concept and dataset.

In summary, prior research has matured along multiple refactoring dimensions—motivation, automation, empirical evolution, and human factors—but agentic refactoring remains unexplored. Our study is the first to provide a large-scale empirical view of agentic refactoring by (i) quantifying its prevalence, (ii) contrasting AI and human refactoring styles, (iii) classifying refactoring intent in AI settings, and (iv) evaluating its quality impact.

## 7 Threats to Validity

In this section, we discuss the threats to validity of our study about agentic refactoring.

### 7.1 Internal Validity

Threats to internal validity concern potential confounding factors that could influence our results. First, our study relies on automated tools for data analysis. RefactoringMiner [1, 57], which we used to identify refactoring operations, is known to produce both false positives and false negatives. Similarly, DesigniteJava [49], used for calculating code quality metrics, has its own inherent limitations. Although these tools are standard in empirical software engineering, their potential inaccuracies could have affected our quantitative results. Second, for RQ3, we employed GPT-4.1-mini to automatically classify the purpose of each refactoring. To mitigate the risk of misclassification, two authors manually labeled a statistically significant subset of the refactorings and we calculated the inter-rater reliability. We achieved a Cohen’s kappa coefficient of 0.77, which indicates a substantial level of agreement. This suggests that while the automated classification may not be perfect, the resulting distribution of purposes is largely reliable.

### 7.2 Construct Validity

Construct validity threats relate to the alignment between our theoretical constructs and our measurements. A primary concern is the definition of an agentic commit. We identify such commits based on keywords and author information from the commit history. However, it is challenging to ascertain the precise extent of human intervention; developers may modify, accept, or reject parts of AI-generated code before committing. To acknowledge this ambiguity, we explicitly frame our study as an analysis of human-AI collaborative refactoring rather than purely autonomous AI contributions.

### 7.3 External Validity

Threats to external validity concern the generalizability of our findings. Our study is based on the AIDev dataset [28], which consists exclusively of open-source software (OSS) projects. The development practices, coding standards, and types of refactoring in industrial, closed-source projects may differ significantly. Furthermore, our analysis was limited to commits that involved changes to Java files. The prevalence and impact of agentic refactoring might vary across different programming languages with distinct ecosystems and tooling support. Therefore, caution should be exercised when generalizing our results to other contexts.

## 8 Conclusion

This study provides the first large-scale empirical analysis of refactoring in agentic software development, examining 15,451 refactoring instances generated by AI agents across real-world open-source Java projects. Our findings clarify the current capabilities, typical uses, and impacts of agentic refactoring.

Our empirical results show that refactoring is a common and intentional activity for AI agents, explicitly targeted in 26.1% of agentic commits. This demonstrates that agents actively participate in software maintenance, frequently engaging in restructuring activities beyond feature implementation or bug fixing. The motivations are overwhelmingly focused on internal code quality: maintainability (52.5%) and readability (28.1%) account for over 80% of cases.

However, our analysis reveals a key limitation. Agentic refactoring is heavily dominated by low-level, consistency-oriented edits such as renaming and type adjustments. Compared to human refactoring, agents perform fewer high-level design changes and more localized modifications, indicating a preference for incremental improvements over architectural restructuring.

Our quantitative assessment shows that agentic refactoring produces statistically significant but small structural improvements, particularly for medium-level changes that combine signature and block modifications. These include measurable reductions in Class Lines of Code (Class LOC median  $\Delta = -15.25$ ) and Weighted Methods per Class (WMC median  $\Delta = -2.07$ ). However, despite the goal of improving quality, agents currently fail to consistently reduce the overall count of known design and implementation smells.

In conclusion, agentic coding tools effectively serve as incremental cleanup partners, excelling at localized refactoring and consistency improvements necessary for long-term maintainability. However, to realize the vision of agents as “software architects,” significant advancements are needed to enable autonomous, architecturally-aware restructuring that consistently addresses higher-level design smells.

## Acknowledgments

We gratefully acknowledge the financial support of JSPS KAKENHI grants (JP24K02921, JP25K21359), as well as JST PRESTO grant (JPMJPR22P3), ASPIRE grant (JPMJAP2415), and AIP Accelerated Program (JPMJCR25U7). We also acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

## References

- [1] Pouria Alikhanifard and Nikolaos Tsantalis. 2025. A Novel Refactoring and Semantic Aware Abstract Syntax Tree Differencing Tool and a Benchmark for Evaluating the Accuracy of Diff Tools. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 34, 2 (2025), 40:1–40:63.
- [2] Eman Abdullah AlOmar, Mohamed Wiem Mkaouer, and Ali Ouni. 2019. Can refactoring be self-affirmed?: an exploratory study on how developers document their refactoring activities in commit messages. In *Proceedings of the 3rd International Workshop on Refactoring (IWOR’19)*. 51–58.

- [3] Eman Abdullah AlOmar, Anushkrishna Venkatakrishnan, Mohamed Wiem Mkaouer, Christian D. Newman, and Ali Ouni. 2024. How to Refactor this Code? An Exploratory Study on Developer-ChatGPT Refactoring Conversations. In *Proceedings of the 21st IEEE/ACM International Conference on Mining Software Repositories (MSR'24)*. 202–206.
- [4] Owura Asare, Meiyappan Nagappan, and N. Asokan. 2023. Is GitHub's Copilot as bad as humans at introducing vulnerabilities in code? *Empirical Software Engineering (EMSE)* 28, 6 (2023), 129.
- [5] Fraol Batole, Abhiram Bellur, Malinda Dilhara, Mohammed Ullah, Yaroslav Zharov, Timofey Bryksin, Kai Ishikawa, Haifeng Chen, Masaharu Morimoto, Shota Motoura, Takeo Hosomi, Tien Nguyen, Hridesh Rajan, Nikolaos Tsantalis, and Danny Dig. 2025. Leveraging LLMs, IDEs, and Semantic Embeddings for Automated Move Method Refactoring. In *Proceedings of the 41th IEEE International Conference on Software Maintenance and Evolution (ICSME'25)*.
- [6] Gabriele Bavota, Bernardino De Carluccio, Andrea De Lucia, Massimiliano Di Penta, Rocco Oliveto, and Orazio Strollo. 2012. When Does a Refactoring Induce Bugs? An Empirical Study. In *Proceedings of the 12th IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM'12)*. 104–113.
- [7] Gabriele Bavota, Andrea De Lucia, Massimiliano Di Penta, Rocco Oliveto, and Fabio Palomba. 2015. An experimental investigation on the innate relationship between quality and refactoring. *Journal of Systems and Software* 107 (2015), 1–14.
- [8] Diego Cedrim, Alessandro Garcia, Melina Mongiovì, Rohit Gheyi, Leonardo da Silva Sousa, Rafael Maiani de Mello, Baldoim Fonseca, Márcio Ribeiro, and Alexander Chávez. 2017. Understanding the impact of refactoring on smells: a longitudinal study of 23 software projects. In *Proceedings of the 11th Joint Meeting on Foundations of Software Engineering (FSE'17)*. 465–475.
- [9] Alexander Chávez, Isabella Ferreira, Eduardo Fernandes, Diego Cedrim, and Alessandro Garcia. 2017. How does refactoring affect internal quality attributes?: A multi-project study. In *Proceedings of the 31st Brazilian Symposium on Software Engineering (SBES'17)*. 74–83.
- [10] Shyam R. Chidamber and Chris F. Kemerer. 1994. A Metrics Suite for Object Oriented Design. *IEEE Transactions on Software Engineering (TSE)* 20, 6 (1994), 476–493.
- [11] Moataz Chouchen, Narjees Bessghaier, Mahi Begoug, Ali Ouni, Eman Abdullah AlOmar, and Mohamed Wiem Mkaouer. 2024. How Do Software Developers Use ChatGPT? An Exploratory Study on GitHub Pull Requests. In *Proceedings of the 21st IEEE/ACM International Conference on Mining Software Repositories (MSR'24)*. 212–216.
- [12] Jonathan Cordeiro, Shayan Noei, and Ying Zou. 2024. An Empirical Study on the Code Refactoring Capability of Large Language Models. *CoRR* abs/2411.02320 (2024).
- [13] Arghavan Moradi Dakhel, Vahid Majdinasab, Amin Nikanjam, Foutse Khomh, Michel C. Desmarais, and Zhen Ming (Jack) Jiang. 2023. GitHub Copilot AI pair programmer: Asset or Liability? *Journal of Systems and Software (JSS)* 203 (2023), 111734.
- [14] Jehad Al Dallal and Anas Abdin. 2018. Empirical Evaluation of the Impact of Object-Oriented Code Refactoring on Quality Attributes: A Systematic Literature Review. *IEEE Transactions on Software Engineering* 44, 1 (2018), 44–69.
- [15] W.W. Daniel. 1990. *Applied Nonparametric Statistics*. PWS-KENT Pub.
- [16] Soham Deo, Divya Hinge, Omkar Sandip Chavan, Yaxuan Olivia Wang, and Mohamed Wiem Mkaouer. 2024. Analyzing Developer-ChatGPT Conversations for Software Refactoring: An Exploratory Study. In *Proceedings of the 21st IEEE/ACM International Conference on Mining Software Repositories (MSR'24)*. 207–211.
- [17] Kayla Depalma, Izabel Miminoshvili, Chiara Henselder, Kate Moss, and Eman Abdullah AlOmar. 2024. Exploring ChatGPT's code refactoring capabilities: An empirical study. *Expert Systems with Applications* 249 (2024), 123602.
- [18] Khaled El Emam. 1999. Benchmarking Kappa: Interrater Agreement in Software Process Assessments. *Empirical Software Engineering* 4, 2 (1999), 113–133.
- [19] Martin Fowler. 1999. *Refactoring - Improving the Design of Existing Code*. Addison-Wesley.
- [20] Deborah H Glueck, Jan Mandel, Anis Karimpour-Fard, Lawrence Hunter, and Keith E Muller. 2008. Exact Calculations of Average Power for the Benjamini-Hochberg Procedure. *The International Journal of Biostatistics* 4, 1 (2008).
- [21] Ahmed E. Hassan, Hao Li, Dayi Lin, Bram Adams, Tse-Hsun Chen, Yutaro Kashiwa, and Dong Qiu. 2025. Agentic Software Engineering: Foundational Pillars and a Research Roadmap. (2025). arXiv:2509.06216 [cs.SE]
- [22] Kosei Horikawa, Yutaro Kashiwa, Bin Lin, Kenji Fujiwara, and Hajimu Iida. 2025. How Does Test Code Differ From Production Code in Terms of Refactoring? An Empirical Study. In *Proceedings of the 41th IEEE International Conference on Software Maintenance and Evolution (ICSME'25)*.
- [23] Yutaro Kashiwa, Kazuki Shimizu, Bin Lin, Gabriele Bavota, Michele Lanza, Yasutaka Kamei, and Naoyasu Ubayashi. 2021. Does Refactoring Break Tests and to What Extent?. In *Proceedings of the 2021 IEEE International Conference on Software Maintenance and Evolution (ICSME 2021)*. 171–182.
- [24] Miryung Kim, Dongxiang Cai, and Sunghun Kim. 2011. An empirical investigation into the role of API-level refactorings during software evolution. In *Proceedings of the 33rd International Conference on Software Engineering (ICSE'11)*. 151–160.
- [25] Miryung Kim, Thomas Zimmermann, and Nachiappan Nagappan. 2012. A field study of refactoring challenges and benefits. In *Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering*

- (FSE'12). 50.
- [26] Miryung Kim, Thomas Zimmermann, and Nachiappan Nagappan. 2014. An Empirical Study of Refactoring Challenges and Benefits at Microsoft. *IEEE Transactions on Software Engineering (TSE)* 40, 7 (2014), 633–649.
  - [27] Noah Lambaria and Tomás Cerný. 2022. A Data Analysis Study of Code Smells within Java Repositories. In *Proceedings of the 17th Conference on Computer Science and Intelligence Systems*, Vol. 32. 313–318.
  - [28] Hao Li, Haoxiang Zhang, and Ahmed E. Hassan. 2025. The Rise of AI Teammates in Software Engineering (SE) 3.0: How Autonomous Coding Agents Are Reshaping Software Engineering. *CoRR* abs/2507.15003 (2025).
  - [29] Jeffrey D. Long, Du Feng, and Norman Cliff. 2003. Ordinal Analysis of Behavioral Data. In *Handbook of Psychology*. John Wiley & Sons, Inc., Hoboken, NJ, USA, Chapter 25, 635–661.
  - [30] H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics* 18 (1947), 50–60.
  - [31] Davood Mazinanian, Nikolaos Tsantalis, Raphael Stein, and Zackary Valenta. 2016. JDeodorant: clone refactoring. In *ICSE (Companion Volume)*. 613–616.
  - [32] Mohamed Wiem Mkaouer, Marouane Kessentini, Slim Bechikh, Kalyanmoy Deb, and Mel Ó Cinnéide. 2014. Recommendation system for software refactoring using innovation and interactive dynamic optimization. In *Proceedings of the ACM/IEEE International Conference on Automated Software Engineering (ASE'14)*. ACM, 331–336.
  - [33] Raimund Moser, Pekka Abrahamsson, Witold Pedrycz, Alberto Sillitti, and Giancarlo Succi. 2007. A Case Study on the Impact of Refactoring on Quality and Productivity in an Agile Team. In *Proceedings of the 2nd IFIP Central and East European Conference on Software Engineering Techniques (Lecture Notes in Computer Science, Vol. 5082)*. Springer, 252–266.
  - [34] Emerson R. Murphy-Hill, Chris Parnin, and Andrew P. Black. 2012. How We Refactor, and How We Know It. *IEEE Transactions on Software Engineering* 38, 1 (2012), 5–18.
  - [35] Mark Kent O'Keeffe and Mel Ó Cinnéide. 2008. Search-based refactoring for software maintenance. *Journal of Systems and Software (JSS)* 81, 4 (2008), 502–516.
  - [36] William F. Opdyke. 1992. *Refactoring object-oriented frameworks*. Ph.D. Dissertation.
  - [37] Fabio Palomba, Andy Zaidman, Rocco Oliveto, and Andrea De Lucia. 2017. An exploratory study on the relationship between changes and refactoring. In *Proceedings of the 25th International Conference on Program Comprehension (ICPC'17)*. IEEE Computer Society, 176–185.
  - [38] Jevgenija Pantiuchina, Bin Lin, Fiorella Zampetti, Massimiliano Di Penta, Michele Lanza, and Gabriele Bavota. 2022. Why Do Developers Reject Refactorings in Open-Source Projects? *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31, 2 (2022), 23:1–23:23.
  - [39] Jevgenija Pantiuchina, Fiorella Zampetti, Simone Scalabrino, Valentina Piantadosi, Rocco Oliveto, Gabriele Bavota, and Massimiliano Di Penta. 2020. Why Developers Refactor Source Code: A Mining-based Study. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 29, 4 (2020), 29:1–29:30.
  - [40] Napol Rachatasumrit and Miryung Kim. 2012. An empirical investigation into the impact of refactoring on regression testing. In *Proceedings of the 28th IEEE International Conference on Software Maintenance (ICSME'12)*. 357–366.
  - [41] Tajimir Rahman, Yuecai Zhu, Lamyea Maha, Chanchal Roy, Banani Roy, and Kevin A. Schneider. 2024. Take Loads Off Your Developers: Automated User Story Generation using Large Language Model. In *Proceedings of the 40th IEEE International Conference on Software Maintenance and Evolution (ICSME'24)*. 791–801.
  - [42] Jeanine Romano, Jeffrey D Kromrey, Jesse Coraggio, Jeff Skowronek, and Linda Devine. 2006. Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen'sd indices the most appropriate choices. In *Proceedings of the 2006 annual meeting of the Southern Association for Institutional Research*. Citeseer, 1–51.
  - [43] Gustavo Sandoval, Hammond Pearce, Teo Nys, Ramesh Karri, Siddharth Garg, and Brendan Dolan-Gavitt. 2023. Lost at C: A User Study on the Security Implications of Large Language Model Code Assistants. In *Proceedings of the USENIX Security Symposium 2023*. 2205–2222.
  - [44] Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. 2025. Vibe Coding vs. Agentic Coding: Fundamentals and Practical Implications of Agentic AI. *CoRR* abs/2505.19443 (2025).
  - [45] Giulia Sellitto, Emanuele Iannone, Zadia Codabux, Valentina Lenarduzzi, Andrea De Lucia, Fabio Palomba, and Filomena Ferrucci. 2022. Toward Understanding the Impact of Refactoring on Program Comprehension. In *Proceedings of the IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER'22)*. 731–742.
  - [46] Agnia Sergeyuk, Yaroslav Golubev, Timofey Bryksin, and Iftekhar Ahmed. 2025. Using AI-based coding assistants in practice: State of affairs, perceptions, and ways forward. *Information and Software Technology (IST)* 178 (2025), 107610.
  - [47] Harsh Mukeshkumar Shah, Qurram Zaheer Syed, Bharatwaaj Shankaranarayanan, Indranil Palit, Arshdeep Singh, Kavya Raval, Kishan Savaliya, and Tushar Sharma. 2023. Mining and Fusing Productivity Metrics with Code Quality Information at Scale. In *Proceedings of the 2023 IEEE International Conference on Software Maintenance and Evolution (ICSME'23)*. 563–567.

- [48] Mahnoosh Shahidi, Mehrdad Ashtiani, and Morteza Zakeri Nasrabadi. 2022. An automated extract method refactoring approach to correct the long method code smell. *Journal of Systems and Software (JSS)* 187 (2022), 111221.
- [49] Tushar Sharma. 2024. Multi-faceted Code Smell Detection at Scale using DesigniteJava 2.0. In *Proceedings of the 21st IEEE/ACM International Conference on Mining Software Repositories (MSR'24)*. 284–288.
- [50] Atsushi Shirafuji, Yusuke Oda, Jun Suzuki, Makoto Morishita, and Yutaka Watanobe. 2023. Refactoring Programs Using Large Language Models with Few-Shot Examples. In *Proceedings of the 30th Asia-Pacific Software Engineering Conference, (APSEC'23)*. 151–160.
- [51] Danilo Silva, João Paulo da Silva, Gustavo Jansen de Souza Santos, Ricardo Terra, and Marco Túlio Valente. 2021. RefDiff 2.0: A Multi-Language Refactoring Detection Tool. *IEEE Transactions on Software Engineering* 47, 12 (2021), 2786–2802.
- [52] Danilo Silva, Nikolaos Tsantalis, and Marco Túlio Valente. 2016. Why we refactor? confessions of GitHub contributors. In *Proceedings of the 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE'16)*. 858–870.
- [53] Alexey Svyatkovskiy, Shao Kun Deng, Shengyu Fu, and Neel Sundaresan. 2020. IntelliCode compose: code generation using transformer. In *Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (FSE'20)*. 1433–1443.
- [54] Gábor Szoke, Gábor Antal, Csaba Nagy, Rudolf Ferenc, and Tibor Gyimóthy. 2014. Bulk Fixing Coding Issues and Its Effects on Software Quality: Is It Worth Refactoring?. In *Proceedings of the 14th IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM'14)*. 95–104.
- [55] Nikolaos Tsantalis, Theodoros Chaikalis, and Alexander Chatzigeorgiou. 2018. Ten years of JDeodorant: Lessons learned from the hunt for smells. In *Proceedings of the 25th International Conference on Software Analysis, Evolution and Reengineering (SANER'18)*. 4–14.
- [56] Nikolaos Tsantalis and Alexander Chatzigeorgiou. 2009. Identification of Move Method Refactoring Opportunities. *IEEE Transactions on Software Engineering (TSE)* 35, 3 (2009), 347–367.
- [57] Nikolaos Tsantalis, Ameya Ketkar, and Danny Dig. 2022. RefactoringMiner 2.0. *IEEE Transactions on Software Engineering* 48, 3 (2022), 930–950.
- [58] Rosalia Tufano, Antonio Mastropaoletti, Federica Pepe, Ozren Dabic, Massimiliano Di Penta, and Gabriele Bavota. 2024. Unveiling ChatGPT's Usage in Open Source Projects: A Mining-based Study. In *Proceedings of the 21st IEEE/ACM International Conference on Mining Software Repositories (MSR'24)*. 571–583.
- [59] Carmine Vassallo, Giovanni Grano, Fabio Palomba, Harald C. Gall, and Alberto Bacchelli. 2019. A large-scale empirical exploration on refactoring activities in open source software projects. *Science of Computer Programming* 180 (2019), 1–15.
- [60] Miku Watanabe, Yutaro Kashiwa, Bin Lin, Toshiki Hirao, Ken-ichi Yamaguchi, and Hajimu Iida. 2024. On the Use of ChatGPT for Code Review: Do Developers Like Reviews By ChatGPT?. In *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering (EASE'24)*. 375–380.
- [61] Miku Watanabe, Hao Li, Yutaro Kashiwa, Brittany Reid, Hajimu Iida, and Ahmed E. Hassan. 2025. On the Use of Agentic Coding: An Empirical Study of Pull Requests on GitHub. *CoRR* abs/2509.14745 (2025).
- [62] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83.
- [63] Dirk Wilking, Umar Farooq Kahn, and Stefan Kowalewski. 2007. An Empirical Evaluation of Refactoring. *e-Informatica Software Engineering Journal* 1, 1 (2007), 27–42.
- [64] Zhenchang Xing and Eleni Stroulia. 2006. Refactoring Practice: How it is and How it Should be Supported - An Eclipse Case Study. In *Proceedings of the 22nd IEEE International Conference on Software Maintenance (ICSM'06)*. 458–468.