

Automated Unit Test Refactoring

YI GAO, Zhejiang University, China

XING HU, Zhejiang University, China

XIAOHU YANG, Zhejiang University, China

XIN XIA, Zhejiang University, China

Test smells arise from poor design practices and insufficient domain knowledge, which can lower the quality of test code and make it harder to maintain and update. Manually refactoring of test smells is time-consuming and error-prone, highlighting the necessity for automated approaches. Current rule-based refactoring methods often struggle in scenarios not covered by predefined rules and lack the flexibility needed to handle diverse cases effectively. In this paper, we propose a novel approach called **UTREFACTOR**, a context-enhanced, LLM-based framework for automatic test refactoring in Java projects. **UTREFACTOR** extracts relevant context from test code and leverages an external knowledge base that includes test smell definitions, descriptions, and DSL-based refactoring rules. By simulating the manual refactoring process through a chain-of-thought approach, **UTREFACTOR** guides the LLM to eliminate test smells in a step-by-step process, ensuring both accuracy and consistency throughout the refactoring. Additionally, we implement a checkpoint mechanism to facilitate comprehensive refactoring, particularly when multiple smells are present. We evaluate **UTREFACTOR** on 879 tests from six open-source Java projects, reducing the number of test smells from 2,375 to 265, achieving an 89% reduction. **UTREFACTOR** outperforms direct LLM-based refactoring methods by 61.82% in smell elimination and significantly surpasses the performance of a rule-based test smell refactoring tool. Our results demonstrate the effectiveness of **UTREFACTOR** in enhancing test code quality while minimizing manual involvement.

CCS Concepts: • **Software and its engineering** → **Automatic programming**; **Software maintenance tools**.

Additional Key Words and Phrases: Test Smells, Test Refactoring, Large Language Models

ACM Reference Format:

Yi Gao, Xing Hu, Xiaohu Yang, and Xin Xia. 2025. Automated Unit Test Refactoring. *Proc. ACM Softw. Eng.* 2, FSE, Article FSE033 (July 2025), 21 pages. <https://doi.org/10.1145/3715750>

1 Introduction

During the software testing process, test code often suffers from **test smells**, which originates from a lack of sufficient domain knowledge by software engineers and the adoption of poor design practices when writing test code. Recent studies [21, 33] have shown that developers tend to prioritize production code over test code, which further contributes to the decline in test code quality. From a software maintenance perspective, test smells in a software system can complicate the test code, making it harder to read, understand, and update. This complexity can ultimately lead to a decline in software product quality and reduced developer productivity. Test smells can be removed through test refactoring—a process of improving the internal structure of the code without altering the software’s external behavior [23, 33, 36]. Numerous empirical studies [21, 26, 29, 32, 33, 36, 37] have highlighted the importance of test refactoring, and developers widely accept the refactored

Authors’ Contact Information: Yi Gao, Zhejiang University, Hangzhou, China, gaoyi01@zju.edu.cn; Xing Hu, Zhejiang University, Hangzhou, China, xinghu@zju.edu.cn; Xiaohu Yang, Zhejiang University, Hangzhou, China, yangxh@zju.edu.cn; Xin Xia, Zhejiang University, Hangzhou, China, xin.xia@acm.org.



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2025 Copyright held by the owner/author(s).

ACM 2994-970X/2025/7-ARTFSE033

<https://doi.org/10.1145/3715750>

tests, and both developers and testers are increasingly recognizing the negative impact of test smells. They widely agree that test quality improves significantly once test smells are eliminated. Additionally, many studies [18, 19, 22, 25, 30, 31, 34, 38–40] have proposed test smell detection tools across programming languages such as Java, Scala, and Python, while also evaluating the impact of test smells on the software development process.

However, manual refactoring of test code remains time-consuming, inefficient, and is prone to errors, underscoring the need for automated tools to address a broader range of test smells. Although there is a clear demand for such tools, only a few open-source options exist for automatic test refactoring [23]. Those tools primarily handle basic and limited types of test smells, leaving many types unaddressed. At present, no tool is capable of efficiently and automatically eliminating all types of test smells.

Given the remarkable capabilities of large language models (LLMs) in understanding, generating, and reviewing code, they might have the potential to play a key role in test refactoring tasks. In this paper, we explore the potential of open-source LLMs to automate unit test refactoring and propose an approach for automatically eliminating test smells in software projects. However, several challenges must be addressed to achieve this goal:

Challenge 1: How to guide the LLM to eliminate test smells in an expected way? Refactoring aims to optimize code structure without altering the original functionality and logic. However, LLMs may generate hallucinations during the refactoring process, producing random code segments that introduce syntax or semantic errors. Thus, the first challenge is how to direct the LLM to follow predefined steps for refactoring, ensuring it remains consistent and produces the intended results.

Challenge 2: How to eliminate multiple test smells simultaneously? Since there are many types of test smells, and multiple smells may exist within a single test method, different removal orders can lead to different refactoring outcomes. The challenge lies in ensuring that all identified test smells are accurately and comprehensively eliminated.

In this paper, we propose a context-enhanced, LLM-based automatic test refactoring approach for Java projects. Our proposed approach, named **UTREFACTOR**, first extracts the project's test code and related context information, such as the methods and classes under test. We then build an external knowledge base that supports test smell elimination, which includes test smell definitions and descriptions, DSL-based refactoring rules, and other relevant contextual information. Next, we simulate the manual refactoring process typically followed by developers or testers. Using a chain-of-thought approach, we guide the LLM through understanding the test's intent, identifying the test smells, and following the DSL-defined refactoring steps to refactor the test code. Additionally, we design a checkpoint mechanism to ensure more thorough smell elimination, particularly when multiple test smells are present in a single test.

We evaluate **UTREFACTOR** on six popular Java open-source projects collected from GitHub. In our experiments, we detect 879 tests with smells out of 9,149 tests and apply smell elimination refactoring to them. The results show that **UTREFACTOR** reduces the number of test smells from 2,375 to 265, achieving a reduction rate of 89%. This represents a 61.82% improvement in performance compared to directly using an LLM for test refactoring. The main contributions of this paper are as follows:

- We propose a novel approach that leverages code context information and integrates refactoring DSL rules to enhance the test refactoring capabilities of LLMs, enabling the automatic elimination of test smells and improving the quality of unit tests.

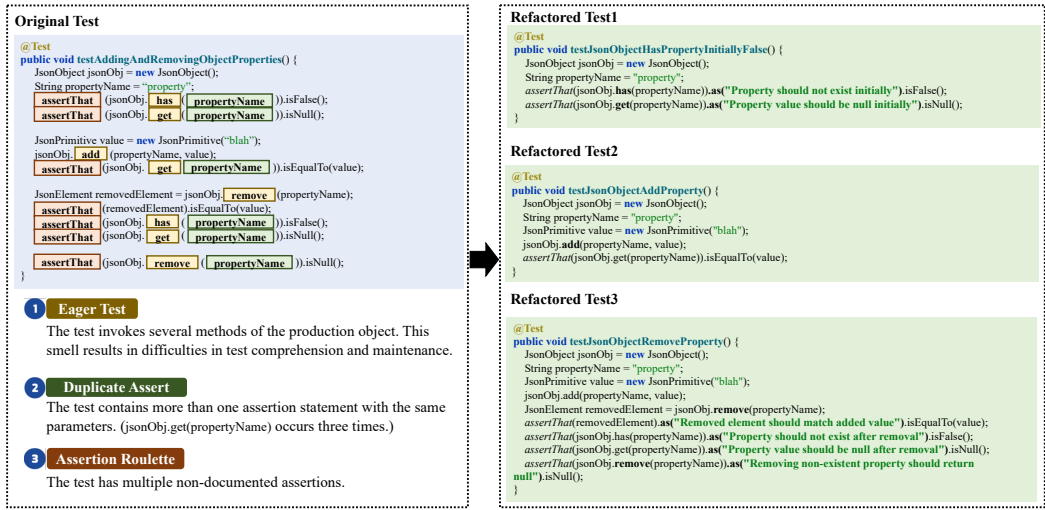


Fig. 1. An example of test refactoring within the Gson.

- We develop UTREFACTOR, a tool that assists developers and testers in automatically refactoring test code in Java projects. It supports various levels of granularity, including single tests, test files, and entire projects.
- We evaluate UTREFACTOR on 879 tests across six open-source projects, reducing the number of test smells from 2,375 to 265, outperforming baselines in effectiveness.

2 Motivation

Developers often prioritize maintaining production code while neglecting the maintenance of test code, which can lead to increased maintenance costs, such as those associated with regression testing. [33]. To illustrate the quality issues in test code, we analyze the Gson project [8], a popular Java library developed by Google, which has over 23.2k stars on GitHub and continues to be actively maintained, including its test code. Figure 1 presents an intuitive example from the Gson's *JsonObjectTest* class. According to the test smell types defined by *tsDetect* [15], this test has been identified with three distinct smells:

❶ Eager Test: This is a common and challenging type of smell to refactor. It occurs when a test method invokes several methods from the production code, making it harder to understand, maintain, and modify the test code effectively, as the test's purpose becomes less clear. In this example, this test suffers from the *Eager Test* smell because it is testing multiple behaviors—adding(add), removing(remove), and checking(has and get) properties—within a single test method. This violates the principle of *Single Responsibility Principle (SRP)* [14]. Each test method should focus on testing a specific behavior or function. Otherwise, it becomes harder to pinpoint the root cause of failures and reduce test clarity. To address this, each of these behaviors should be split into separate test methods to ensure that the test remains focused and maintainable.

❷ Duplicate Assert: In this example, the repeated use of the same parameter *propertyName* across multiple assertions leads to the identification of a *Duplicate Assert* smell. Eliminating this smell can be achieved by either splitting the original test or transforming it into a *Parameterized Test*, which removes the need for duplicate assertions by declaring parameterized values in an annotation and executing the test method multiple times.

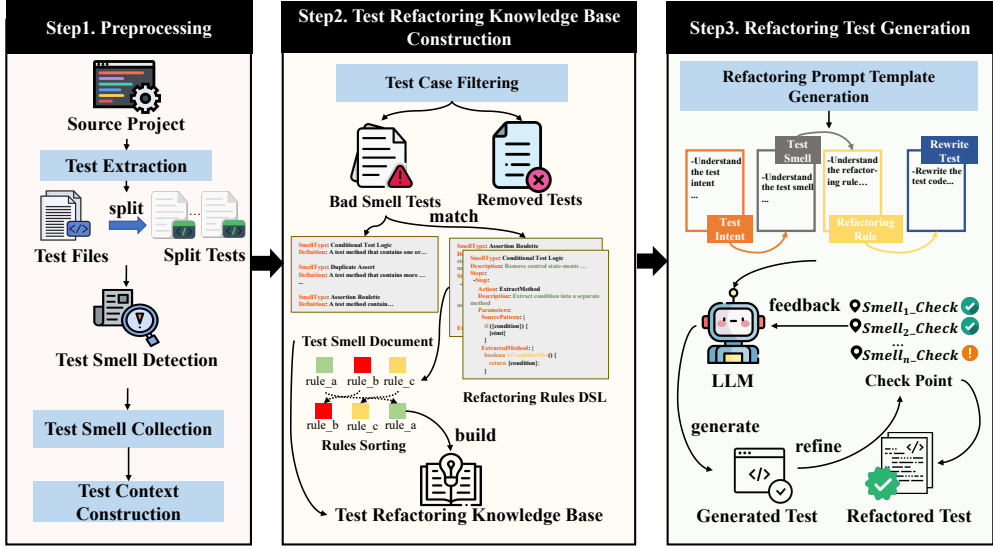


Fig. 2. Overview of our approach.

③ **Assertion Roulette:** This is another common smell, indicating that multiple assertions in a test lack clear identification of the reasons for failure, and make it difficult to locate the failing assertion. To address this, each assertion should be supplemented with descriptive information that clarifies its purpose. This allows developers to quickly identify and fix issues when the test fails.

These test smells demonstrate the potential risks in test code that can compromise its quality and maintainability, highlighting the importance of systematically eliminating the existing smells. Our tool, UTREFACTOR, is designed to automatically refactor unit tests and eliminate test smells, thereby improving the quality of test code in existing software projects. Figure 1(right) shows tests refactored using UTREFACTOR, which have successfully removed the three identified test smells. Developers or testers can use UTREFACTOR to refactor individual unit tests, single test files, or all tests in a project at once to eliminate smells and improve the overall quality of the software's code.

Another potential benefit of automated test refactoring lies in its impact on LLMs, many of which are trained on data from open-source communities. By improving the quality of the test code that serves as training data, we can enhance the performance of LLMs in tasks such as test generation. For instance, during the data preprocessing phase, we can refactor test code that contains smells and use the cleaned, refactored code as training data. This approach may lead to more effective and reliable LLM-generated tests.

3 Approach

This section presents the details of our proposed approach UTREFACTOR. As shown in Figure 2, it can be divided into three main steps:

Step ① Preprocessing. We begin by extracting the test code from the project and detecting all test smells present. For tests that need to be refactored due to the presence of smells, we also extract relevant code context (such as the tested methods and classes) during this step, treating this context as external knowledge for the LLM.

Step ② Test Refactoring Knowledge Base Construction. In this step, we design the types and definitions of test smells to serve as external knowledge that the LLM can reference. Additionally,

Table 1. Extracted test-related context information from the project.

Items	Description
Package Name	The name of the package where the focal method is located, eg., <code>org.jsoup.parser</code>
Focal Class	The name of the class containing the focal method. eg., <code>public class Parser</code>
Focal Method Signature	The signature of the focal method. eg., <code>public static Document parse(String html, String baseUrl)</code>
Focal Method Comment	The code comment for the focal method. eg., <code>Parse HTML into a Document.</code>
Other Invoke Methods	The signatures of other methods invoked in the test. eg., <code>public static String unescapeEntities(String string, boolean inAttribute).</code>

we use a Domain-Specific Language (DSL) [28] to define corresponding refactoring rules for each type of test smell, constructing a comprehensive refactoring knowledge base.

Step ③ Refactored Test Generation. This step focuses on generating the refactored test code. We design a specialized prompt template tailored for the test refactoring task. Moreover, we employ the Chain-of-Thought (CoT) [6] reasoning approach and propose a checkpoint mechanism to improve the LLM’s performance in refactoring test code.

3.1 Preprocessing

As shown in Figure 2, this step involves three key processes: extracting the test code, identifying test smells, and gathering relevant contextual information about the tests.

3.1.1 Test Extraction. First, UTREFACTOR automatically gathers all test files from the test directory (e.g., `src/test`) within a given Java project. To avoid unnecessary detection of code smells and refactoring overhead, during the extraction process, we analyze whether the test files contain at least one method annotated with `@Test` and automatically filter out files that do not have the `@Test` annotation. Besides, to analyze the focal methods in a fine-grained approach and meet the requirements of test smell detection, we collect the corresponding classes under test in this step. Specifically, we strip the prefix and suffix `Test` from the test file names (e.g., `ParserTest`→`Parser`), and use this as an index to search for matching class files in the `src/main` directory. If a match is successful, the test file and the corresponding class under test are paired and collected together.

3.1.2 Test Smell Detection. To detect test smells present in the project, we integrate the *tsDetect* [31], an automated test smell detection tool for Java projects.

Refining Detection Granularity. Our UTREFACTOR performs refactoring at the level of individual test methods, refactoring one test at a time. However, the *tsDetect* operates at a file-level granularity, meaning it can detect test smells in a test file as a whole but cannot pinpoint smells within specific test methods. To support the detection of smells at the level of individual test methods, we refine the test smell detection granularity. Specifically, we employ a split-and-merge approach: before smell detection, we split each test file into multiple sub-files, with each sub-file containing a single test method and its necessary context. This ensures that *tsDetect* can analyze

each test method independently. After the refactoring process is completed, these sub-files are merged back together into a complete test file based on the original split index.

3.1.3 Test Context Collection. Since refactoring operations are aimed at optimizing the code structure without altering its functional logic, it is crucial to ensure that the LLM has a sufficient understanding of the test code's functionality before refactoring. Moreover, Yuan et al. [41] have demonstrated that providing additional test context can significantly enhance the performance of LLMs in test code generation tasks. Inspired by these findings and applying them to the task of refactoring test code, we extract relevant contextual information from tests that exhibit smells and use it as external knowledge to help the LLM better comprehend the original test's intent, as shown in Table 1.

In this step, we extract the contextual information for each test that requires refactoring. During the subsequent refactoring steps, this information is integrated into the refactoring prompt template as a knowledge source to assist the LLM in accurately understanding the test's original intent.

3.2 Test Refactoring Knowledge Base Construction

In this step, we design the test smell types and corresponding refactoring rules as applicable external knowledge, aiming to enhance the LLM's understanding of test smells and effectively eliminate those present in unit tests.

3.2.1 Test Smell Knowledge. In the smell detection step, we integrate *tsDetect*, which uses the 19 types of test smells defined by Peruma et al. [31]. Although they define test smells and provide relevant code examples, the definition of test smells is not standardized across existing research, and the types and numbers of smells can vary between studies.

Additionally, LLMs are trained on data from various sources, such as blogs and GitHub repositories, which can lead to significant differences between LLMs in terms of training data. This variability leads to inconsistencies in how LLMs interpret test smells, which in turn affects the effectiveness of the refactoring process. To illustrate this issue, we use *LLaMA-70b* [12]. As shown in Figure 3, when asked about the meaning of specific test smells, the LLM's responses deviate from our expectations. For example, *tsDetect* does not include **Fragile Test** in its list of smells. Additionally, for the **Mystery Guest** smell,

LLaMA-70b provides a vague explanation, whereas *tsDetect* defines it as the presence of unused variables in the *setup/teardown* method. There are also instances where the definitions are similar, but the naming conventions differ. For example, **Missing Assertion** refers to a test method that lacks a test oracle, but in *tsDetect*, this is categorized as **Unknown Test**.

To mitigate the impact of these inconsistencies on the refactoring process, we provide the LLM with a clear and standardized set of test smell definitions before refactoring. Given that Peruma et al. [31] provide comprehensive definitions, explanations, and code examples for 19 test smells in Java, we use these to build part of the external knowledge base for test smells, providing it to the LLM to eliminate inconsistencies in smell interpretation that could impact the refactoring process. Due to space limitations, the detailed definitions of these smells are available in our replication package [16]. For each smell type, we provide a structured definition, including a description of the

Question

As an expert in the field of **test code refactoring**, please enumerate the **types of test code smells** along with their definitions.

Llama3-70B Output

- **Missing Assertion**
A test that lacks assertions, making it impossible to verify the expected behavior.
- **Rigidity Test**
A test that is **difficult** to modify or extend, making it hard to adapt to changing requirements or new functionality.
...
- **Mystery Guest**
A test that uses **complex** or **unclear** setup/teardown, making it hard to understand the test's purpose.

Fig. 3. An example of test smell explanation of the LLM.

test smell, its impact, and a pseudocode example to illustrate the concept. During the refactoring process, these standardized smell definitions guide the LLM's interpretation of the smells present in the test, ensuring consistent and accurate refactoring results.

3.2.2 Test Refactoring Rule DSL. To eliminate the smells present in tests, we design corresponding refactoring rules for each type of smell. We draw on previous manual approaches to eliminating test smells and empirical research on test smell refactoring [17]. Based on these insights and best testing practices, we formalize DSL rules for each type of test smell, which form the foundation of the entire refactoring process. In designing these DSLs, we exclude simple smell types such as *Default Test*, which refers to default test classes automatically created by *Android Studio* when a project is initialized and do not require refactoring. Similarly, we handle *Ignored Test* (tests marked as ignored) and *Empty Test* (tests with empty method bodies) by simply removing the unit tests that exhibit these smells. Finally, we define 13 method-level DSLs for Java test smells, aimed at eliminating smells in unit tests. Due to space limitations, the complete set of DSLs is available in our replication package [16]. Since these DSLs determine the methods and steps for eliminating test smells, which is crucial for unit test refactoring, we evaluate the correctness of the DSL design by assessing whether the refactored tests successfully eliminate the smells without introducing new issues. This will be discussed in detail in subsequent sections.

We leverage LLMs to refactor tests with identified smells according to the provided refactoring rules. A common approach is to describe these refactoring rules in natural language as prompts for the LLM. While this approach is straightforward and easy to understand, it presents challenges in practice: natural language can be ambiguous or imprecise, which may lead the LLM to generate code that does not meet expectations, especially when multiple smells are present in a single test. Without a clear strategy to guide the refactoring process, the LLM's behavior can become unpredictable, resulting in inconsistent outcomes. Moreover, the lengthy natural language descriptions required for such complex cases can hinder the LLM's effectiveness in refactoring tests.

To address this issue, we propose using a more precise Domain-Specific Language (DSL) to express the test smell refactoring rules. The advantage of using a DSL is that it allows for more precise and standardized descriptions of refactoring steps, thereby reducing the risk of misinterpretation. Additionally, DSLs are typically structured, which facilitates the extension and maintenance of new refactoring rules. In practice, developers can easily add and update refactoring rules within the defined DSL structure. This structured approach ensures that refactorings are consistent, predictable, and less prone to errors, providing a clear framework for expanding the tool to accommodate new test smells and evolving best practices in refactoring.

The DSL we design for test smell refactoring is illustrated in Figure 4. Specifically, the *SmellType* defines the specific category of test smell being addressed, ensuring that the appropriate refactoring strategy is applied. The *Description* provides a concise explanation of how the refactoring will

```

RefactorRule ::= {
  SmellType: String,           # Corresponds to the type of test smell defined in
  Description: String,         the tsDetect tool.
  Steps: [Step],
  Example: Example,
  Variable: Variable
}
Step ::= {
  Description: String,
  Action: ActionType,
  Parameters: {ParameterName: ParameterValue}
}
ActionType ::= AddAnnotation | ReplaceMethodCall |
  AddMessageToAssert | ExtractMethod | RemoveConditional |
  ConvertToParameterizedTest | RefactorMethod
ParameterName ::= TargetType | Annotation | TargetMethod |
  ReplacementPattern | Import | SourcePattern | ExtractedMethod |
  MessagePattern
Example ::= {Before: String, After: String}
Variable ::= {VariableName: String, VariableValue: String}

```

Fig. 4. A hierarchical definition of the DSL structure for test refactoring rules.

Assertion Roulette		
SmellType: Assertion Roulette Description: Add descriptive messages to assert statements. Steps: -Step: Action: AddMessageToAssert Description: Add descriptive message to each assert statement. Parameters: MessagePattern: [assertionMessage]	Original Code Example <pre>@Test public void testRoot() { ...// define span, el, el2 assertEquals("<a>Hello", el.outerHtml()); assertNotNull(span); assertEquals(el, el2); }</pre>	Refactored Code Example <pre>@Test public void testRoot() { ...// define span, el, el2 assertEquals("<a>Hello", el.outerHtml(), "Element outerHtml should match expected output."); assertNotNull("Selected span element should not be null.", span); assertEquals("Root of span should be the initial element.", el, el2); }</pre>

(a) The definition of DSL rules for eliminating AR (Assertion Roulette) smell

Magic Number Test		
SmellType: Magic Number Test Description: Extract magic numbers and assign meaningful names. Steps: -Step: Action: IdentifyMagicNumbers Description: Identify magic numbers in assertions. Parameters: SourcePattern: assert({magicNumber}, {variable}) -Step: Action: ExtractVariable Description: Extract number to a variable with meaningful name. Parameters: SourcePattern: assert({magicNumber}, {variable}) ReplacementPattern: int {meaningfulName} = {magicNumber}; assert({meaningfulName}, {variable})	Original Code Example <pre>@Test public void testIteratorRemovable(){ Attributes a = new Attributes(); a.put("Tot", "a&p"); a.put("Hello", "There"); a.put("data-name", "Jsoup"); Iterator<Attribute> iterator = a.iterator(); Attribute attr = iterator.next(); assertEquals("Tot", attr.getKey()); ... iterator.remove(); assertEquals(2, a.size()); }</pre>	Refactored Code Example <pre>@Test public void testIteratorRemovable_hasKeyAfterRemove() { Attributes a = new Attributes(); a.put("Tot", "a&p"); a.put("Hello", "There"); a.put("data-name", "Jsoup"); Iterator<Attribute> iterator = a.iterator(); Attribute attr = iterator.next(); assertEquals("Tot", attr.getKey()); iterator.remove(); int expectedSize = 2; assertEquals(expectedSize, a.size(), "After removing 'Tot', there should be 2 attributes left"); }</pre>

(b) The definition of DSL rules for eliminating MNT (Magic Number Test) smell

Duplicate Assert		
SmellType: Duplicate Assert Description: Eliminate duplicate assertions with parameterized tests. Steps: -Step: Action: ConvertToParameterizedTest Description: Convert test to a parameterized test. Parameters: TargetType: Method -Step: Action: AddAnnotation Description: Declare parameterizable values in @CsvSource annotation. Parameters: TargetType: Method Annotation: @CsvSource({ "{paramValues}" }) -Step: Action: ModifyAssert Description: Use parameterized values. Parameters: SourcePattern: public void T() { {stmts} } ReplacementPattern: public void T({params}) {{refactoredStmts}}	Original Code Example <pre>@Test public void testWholeText() { Document doc = Jsoup.parse("<p> Hello\nthere &nbsp; </p>"); assertEquals(" Hello\nthere ", doc.wholeText()); doc = Jsoup.parse("<p>Hello \n there</p>"); assertEquals("Hello \n there", doc.wholeText()); doc = Jsoup.parse("<p>Hello <div>\n there</div></p>"); assertEquals("Hello \n there", doc.wholeText()); }</pre>	Refactored Code Example <pre>@ParameterizedTest @CsvSource({ "<p> Hello\nthere &nbsp; </p>," "Hello\nthere ", "<p>Hello \n there</p>," "Hello \n there" }) public void testWholeText(String input, String expected) { Document doc = Jsoup.parse(input); assertEquals("The whole text does not match the expected value.", doc.wholeText(), is(expected)); }</pre>

(c) The definition of DSL rules for eliminating DA (Duplicate Assert) smell

Fig. 5. DSL rules for refactoring three types of test smells, accompanied by code examples illustrating the test code before and after refactoring.

be executed for this smell type. The *Steps* component breaks down the refactoring process into a sequence of actions, enabling a fine-grained, step-by-step approach to modifying the test code. Additionally, the *Example* provides before-and-after code snippets to demonstrate the practical application of the refactoring, aiding in both understanding and verification. Finally, *Variables*

parameterize the refactoring logic, offering flexibility and reusability across different test cases. With this structure, the DSL systematically defines and executes complex code refactoring operations, which facilitates the automation of the test refactoring process.

Next, we present specific refactoring rules for test smells. Figure 5 illustrates three concrete examples of these refactorings. For example, when a test contains multiple assert statements without descriptive messages, it is flagged with the *Assertion Roulette* smell. During refactoring, the test matches the DSL rule for the type *Assertion Roulette*. This rule guides the refactoring process by specifying the action *AddMessageToAssert*, which involves adding meaningful descriptive messages to each assert statement. Another example is when a test is identified with the *Duplicate Assert* smell. In this case, the test matches the corresponding DSL rule for *Duplicate Assert*, which outlines a three-step refactoring process to eliminate the smell: **Step ❶** Replace the `@Test` annotation with `@ParameterizedTest` to enable parameterized testing. **Step ❷** Add the `@CsvSource` annotation. This allows additional test cases to be added as parameters in `@CsvSource`, eliminating the need for multiple `assertEquals` statements. **Step ❸** Reduce duplicate code by writing the test logic only once. Parameterized testing makes the test cases more intuitive and clearly expresses the relationship between different inputs and expected outputs.

We put the details of DSLs on our replication package [16]. These rules are then provided to the LLM as external knowledge during the refactoring process, guiding the LLM to refactor the test according to the predefined rules.

3.2.3 Sorting Refactoring Rules. Considering that a single test often exhibits multiple types of smells, it is crucial to determine the appropriate order of refactoring operations, as the sequence can significantly impact the effectiveness of the final test code refactoring. To address this, we categorize the test smell refactoring rules and assign them an execution priority. The LLM is then instructed to eliminate test smells according to their priority, from highest to lowest.

We have classified refactoring operations into three categories based on their characteristics: *Removal*, *Structural Optimization*, and *Functional Optimization*. **❶ Removal:** This category includes smells that require the test to be removed from the refactoring list rather than undergoing structural adjustments. Examples include *EmptyTest* (test method with an empty body), *Unknown Test* (test without assertions), and *Default Test* (automatically generated default tests). Prioritizing removal operations helps to avoid unnecessary refactoring efforts for tests that do not contribute to code quality. **❷ Structural Optimization:** This category focuses on altering the existing structure of the test code. Examples include *Eager Test*, where a test method calls multiple methods of the production object, and *Duplicate Assertion*, where identical assertions are repeated within the same test method. **❸ Functional Optimization:** This category targets improvements to assertion statements, such as *Assertion Roulette* (where multiple assertions lack descriptive messages).

Structural Optimizations take precedence over *Functional Optimization* because a well-structured test provides a solid foundation for functional improvements. For example, the *Assertion Roulette* refactoring rule involves adding descriptive messages to assertions, and the *Magic Number Test* refactoring rule extracts numeric literals into named variables. However, the *Duplicate Assertion* refactoring rule eliminates duplicate assertions using parameterized tests, which could render the optimizations of *Assertion Roulette* unnecessary if performed afterward. Therefore, it is crucial to refactor *Duplicate Assertion* before applying other functional optimizations.

By following this prioritized refactoring approach, we ensure that the test code is improved in an orderly manner, starting with essential removals, followed by structural enhancements, and concluding with functional refinements.

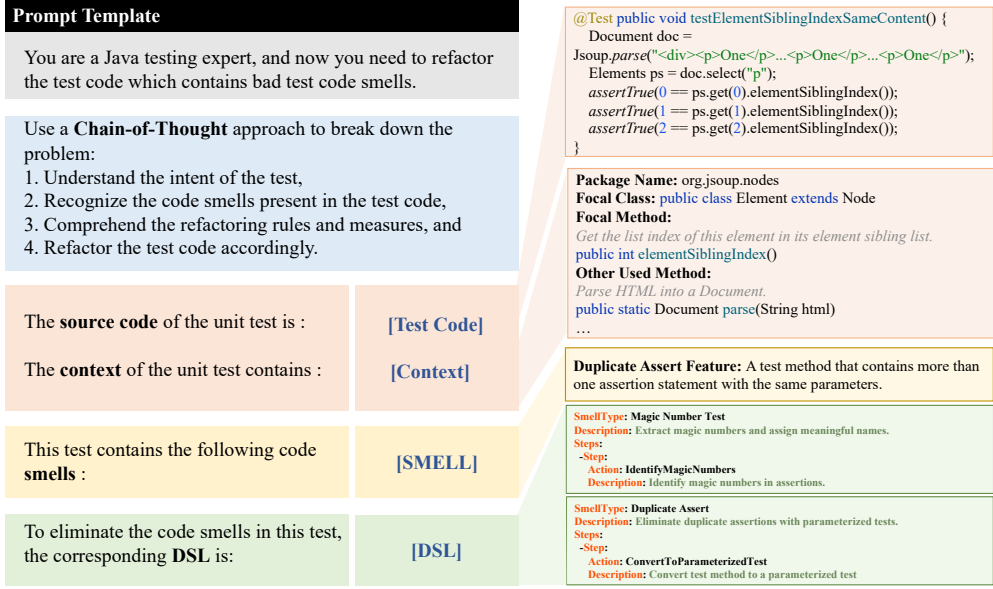


Fig. 6. The Prompt Template used for test smell elimination, along with corresponding examples of prompts.

3.3 Refactoring Test Generation

In this final step, we generate the refactored test code using the specially designed prompt template. The template guides the LLM in applying the refactoring rules to eliminate the identified smells from the test code.

3.3.1 Test Refactoring Prompt Design. Typically, when a developer or tester refactors a test, they follow a process that involves: ❶ Understanding the intent of the test code, including what the focal method does and how the assertions verify its behavior. ❷ Identifying the quality issues in the current test, specifically recognizing any smells present. ❸ Most importantly, determining how to refactor the test to eliminate these smells. ❹ Rewriting the test code accordingly. Our approach is designed around these steps and is divided into four steps. This four-step strategy is based on the Chain-of-Thought (CoT) paradigm, which enhances LLM reasoning capabilities by encouraging step-by-step thinking, leading to better outcomes in tasks such as extraction and reasoning [24].

As illustrated in Figure 6, the complete prompt template based on CoT is designed to improve the LLM's ability to refactor tests. In guiding the LLM through the refactoring process, we simulate the actual steps a developer takes. First, to understand the test's intent, we provide the LLM with the test code along with relevant context, such as the tested class and method signatures. Next, we present the identified smells in the test and their corresponding definitions. Crucially, we supply the DSL-defined refactoring rules tailored to each smell type. Finally, with this external knowledge, the LLM proceeds to refactor the test. This structured approach ensures that the LLM closely follows a logical and thorough process, much like a human developer or tester, resulting in more effective and accurate test refactoring.

3.3.2 Checkpoint Mechanism. When a test contains multiple smells, even with sufficient information provided, the LLM may still miss eliminating some of the smells during refactoring. For example, in the `testSizeWhenHasInternal` method from the Jsoup project, the test has four identified smells: *Assertion Roulette*, *Eager Test*, *Duplicate Assert*, and *Magic Number Test*. Despite explicitly

instructing the LLM to address all four types of smells in the prompt, in practice, the LLM often fails to fully refactor tests with multiple smells.

To address this issue and mitigate incomplete refactoring by the LLM, we design a checkpoint mechanism that allows the LLM to self-verify the completeness of its refactoring process. Specifically, for each smell in the test, we set up a checkpoint that asks whether the refactored test still exhibits the given issue. If the issue persists, the LLM continues refining the refactoring to eliminate the smell. For tests with multiple smells, we construct a chain of smell checkpoints, as illustrated in Figure 2. In the case of the test containing four smells, we collect the corresponding four checkpoints and integrate them into the prompt. The LLM uses this list to systematically check whether the smells have been fully removed. If any issues remain, the LLM will further adjust its refactoring until the test is completely free of the identified smells. This checkpoint mechanism ensures that even in complex cases with multiple smells, the LLM can perform thorough and effective refactoring, ultimately leading to higher-quality test code.

3.3.3 Refactoring Test Generation. It is essential for developers or testers to understand which tests have been refactored and the reasons behind these changes. To address this, we generate a detailed test refactoring report in this step. This report includes a list of refactored tests, specifying the file locations and test method names. It also outlines the detected smells for each test and the corresponding refactoring methods applied. Additionally, to facilitate subsequent testing and compilation, we organize the refactored test files and methods according to the original project's test structure.

4 Evaluation

Our experiments are designed to address the following research questions:

- **RQ1: How is the quality of the test code after the refactoring of `UTREFACTOR`?**
- **RQ2: How effective is the refactoring of `UTREFACTOR`?**
- **RQ3: How effective is `UTREFACTOR` in eliminating each category of Test Smell?**
- **RQ4: What is the time efficiency of `UTREFACTOR` in test refactoring?**

4.1 Experimental Setup

Dataset. We collect projects from GitHub, including those from well-known organizations (e.g., Apache [1]) and popular projects (e.g., Jsoup [11] with 10.9k stars). To facilitate the test refactoring process, we set the following criteria for collection: first, to simplify test smell detection and make refactoring easier to validate, we collect projects written in Java and managed by Maven. Second, the source code of the project must compile successfully, and all unit tests are required to pass. Finally, each project must exhibit at least 100 test smells as detected by the *tsDetect* [15]. Ultimately, we select six open-source projects for our test refactoring experiments. As shown in Table 2, `UTREFACTOR`'s integrated tool *tsDetect* identifies 879 unit tests with smells across these projects, with a combined total of 2,375 test smells detected.

Baselines. For the baseline comparison of refactoring tools, we choose two tools for our unit test refactoring benchmarks. First, *TESTAXE* [23], a tool implemented using the Rascal [13] meta-programming language, primarily designed to automatically upgrade projects from *jUnit 4* to *jUnit 5* to automatically detect and eliminate test smells in Maven-managed Java projects. It is rule-based and leverages features in the *jUnit 5* testing framework to remove test smells. Second, considering that large language models (LLMs) inherently have the capability to refactor test code, we employ a default open-source LLM (*Llama3-70B*) with a generic prompt as another comparison baseline. As depicted in Figure 7, we use a standard prompt template, which provides the *Llama3-70B* with the

Table 2. Test refactoring dataset (#T`Smell` - number of tests with smells, #TS `Num` - number of test smells).

Project	Tests	#T <code>Smell</code>	#TS <code>Num</code>	LOC
Commons-cli [3]	334	85	149	5,716
Commons-compress [4]	1,585	37	263	29,249
Commons-math [5]	2,723	101	196	65,530
Gson [8]	1,348	293	788	21,416
Jfreechart [10]	2,291	148	324	40,354
Jsoup [11]	868	215	655	12,145
Total	9,149	879	2,375	174,510

original test code and all detected test smell types. The LLM is then tasked with removing these smells and generating the refactored test code.

Role: *You are a Java testing expert, and now you need to refactor the test code which contains bad test code smell.*

User: Given the following `Unit Test`, and the `Test Smells` it contains, eliminate the smells present in this test without altering the original functionality and logic of the test, and provide the refactored test code.

Code and Smells: `[Unit Test]` `[Test Smells]`

Fig. 7. Prompt Template for default LLM used to eliminate smell in tests.

4.2 RQ1: How is the quality of the test code after the refactoring of UTREFACTOR?

Since the test refactoring process must eliminate test smells without altering the original test functionality, we first verify the quality of the refactored test code to ensure effective refactoring. This includes checking for code errors and ensuring that the functionality remains unchanged. Specifically, we first assess the syntax correctness of the refactored test code. To achieve this, we calculate the compilation pass rate and test execution pass rate of all refactored tests, evaluating whether UTREFACTOR introduces any errors during the structural changes that lead to compilation or runtime failures.

Table 3 presents the quality of the refactored test code across all open-source projects. As shown, UTREFACTOR refactor 879 unit tests with identified test smells from six open-source projects. The #CPR column represents the compilation pass rate of the refactored tests. We manually compile all refactored tests and calculate that the average compilation pass rate is 95%, with the highest rate being 97% in the *Jsoup* project. #EPR represents the execution pass rate of the refactored tests. Similarly, we manually execute all refactored tests and calculate an average #EPR of 89%, with the highest rate reaching 92% in the *Gson* and *Jsoup* projects. This indicates that while the refactoring process did introduce some errors, it remains largely effective, with the overall results demonstrating a solid level of reliability.

We analyze the tests where UTREFACTOR introduces errors after refactoring. In the *Gson* project, `testStrictComments` contains three assertions (`assertStrictError`) without descriptive information, resulting in the detection of an AR (Assertion Roulette) smell. During the refactoring process, UTREFACTOR generates descriptive messages for each of the three assertions. However, a parameter mismatch error occurs during compilation. Upon further analysis, we find that `assertStrictError`

Table 3. Quality of refactored test code generated by UTREFACTOR (#RTests - number of tests after refactoring, #CPR - test compilation pass rate, #EPR - test execution pass rate).

Project	#RTests	#CPR	#EPR	Line Coverage	Branch Coverage
				cov.(unchgd.)	cov.(unchgd.)
Commons-cli	126	93%	89%	98% (✓)	95% (✓)
Commons-compress	65	92%	83%	84% (✓)	75% (✓)
Commons-math	144	95%	88%	92% (✓)	85% (✓)
Gson	469	94%	92%	87% (✓)	85% (✓)
Jfreechart	314	95%	81%	54% (✓)	45% (✓)
Jsoup	404	97%	92%	86% (✓)	80% (✓)
Total	1,522	95%	89%	77% (✓)	67% (✓)

is a built-in assertion in *Gson*, different from the standard *JUnit* assertions. This assertion only accepts two parameters: an exception describing the expected behavior and an expectedLocation parameter. It does not support an additional message parameter. This conflicts with UTREFACTOR's DSL rule for refactoring AR smells, which requires assertions to include a message parameter, leading to a compilation error and causing the smell to remain.

We calculate the coverage of tests that still pass after refactoring by using the *Jacoco* [9] test coverage tool in each project. As shown in Table 3, we observe that UTREFACTOR does not alter the original line coverage or branch coverage of the tests after refactoring. This indicates that UTREFACTOR only restructures the code, such as refactoring assertions or breaking down test methods, without changing the original functionality. It avoids introducing new calls to the project's production APIs or omitting existing conditions or function calls within the tests. This outcome aligns with our expectations, as the goal of UTREFACTOR is to eliminate test smells and improve test code quality, rather than modifying the underlying functionality.

Consistency of Refactoring. Given the challenge of determining functional and behavioral consistency of test code before and after refactoring, we employ a more advanced model (GPT-4o) to evaluate the functional consistency of test code pre- and post-refactoring, supplemented by a manual secondary review of the evaluation results. Specifically, we provide GPT-4o with all 879 pairs of test code that can be executed successfully before and after refactoring to assess whether the changes involve only structural adjustments while maintaining functional consistency.

Table 4. Classification of test inconsistencies before and after refactoring.

Inconsistent Type	Num
Exception Handling	18
Redundant Assertion	9
Redundant Variable Declaration	5

Table 4 indicates that 32 cases (approximately 4%) are flagged as **Inconsistent**, which can be categorized into three types: Firstly, regarding differences in exception handling (18 cases), although GPT-4o identifies a behavioral change between the original test code's throws exception handling and the DSL-refactored JUnit assertDoesNotThrow method, this refactoring actually aligns with the design philosophy of the JUnit 5 framework, providing a more standardized exception verification mechanism.

Secondly, in terms of redundant assertion handling (9 cases), the current DSL strategy of directly deleting redundant assertions is marked by GPT-4o as a potential behavioral change, prompting us to plan an optimization of the refactoring rules to comment out redundant assertions and add developer prompts. Lastly, concerning redundant variable declarations (5 cases), variables defined during the setup phase are redundantly declared in subtests when splitting the original test into

multiple subtests. Although there is still room for improvement in UTREFACTOR's handling of redundant assertions and duplicate variables, it performs excellently overall in maintaining test functional consistency.

4.3 RQ2: How effective is the refactoring of UTREFACTOR?

We are next interested in evaluating the extent to which UTREFACTOR eliminates bad smells in refactored test code that still passes execution, thus validating the effectiveness of UTREFACTOR in test refactoring. Additionally, we compare and analyze UTREFACTOR's effectiveness against two baseline tools. To assess refactoring effectiveness, we focus on two key metrics in this research question (RQ): the number of passing tests after refactoring and the number of remaining smells in those tests. We also measure the change in the total number of smells before and after refactoring. A higher reduction rate indicates greater effectiveness of our automated refactoring approach in minimizing test smells across the project.

Table 5 shows the effectiveness of UTREFACTOR and the two baseline tools in refactoring tests across six Java open-source projects. Out of the 2,375 test smells present across the six projects, UTREFACTOR reduced the number to 265 after refactoring, achieving an average reduction rate of 89%. The *Commons-compress* project shows the highest reduction rate, reaching 94%. This demonstrates that UTREFACTOR is highly effective in reducing test smells. In particular, for the most prevalent smell, AR (Assertion Roulette), UTREFACTOR had the most significant impact, completely eliminating this smell by successfully adding descriptive messages to all assertions that lacked them. The effectiveness of UTREFACTOR in addressing different types of test smells will be discussed further in RQ3.

Comparison with TESTAXE. As shown in Table 5, TESTAXE achieves a 19% smell reduction rate in the *Commons-math* project, while in the *Commons-compress* and *Jsoup* projects, the reduction rate is less than 1%. In the *Commons-cli*, *Gson*, and *Jfreechart* projects, TESTAXE fails to eliminate any test smells. TESTAXE primarily aims to upgrade projects from JUnit 4 to JUnit 5. During this process, it uses JUnit 5 features to refactor existing tests and eliminate test smells. TESTAXE relies on a set of built-in syntax matching and replacement rules for test refactoring. However, it includes only five rules related to smell elimination, limiting its effectiveness. For example, it cannot handle AR (Assertion Roulette), one of the most common test smells. Additionally, TESTAXE's replacement rules require strict matching with the original test code, meaning that any code not covered by these rules remains unchanged. The biggest limitation of their approach is its poor adaptability. For example, in handling ECT (Exception Catching Throwing) smells, TESTAXE uses the ExpectedExceptionTransformation rule, which only matches patterns like throws in function signatures. However, test cases often involve exceptions handled in try-catch blocks that require refactoring, and since TESTAXE's rules do not cover these patterns, it cannot eliminate such exception smells. These limitations significantly reduce TESTAXE's time efficiency in eliminating and refactoring test smells.

Comparison with Llama3-70B. It is evident that when only providing the test code and corresponding smell types to Llama3-70B for refactoring, it reduces a total of 1,080 smells across the six projects, with an average reduction rate of 55%. This is lower than that achieved by UTREFACTOR. This shows that while the LLM has the ability to refactor test smells, its effectiveness is limited. After manual inspection of the LLM-refactored tests, we find that it performs best in handling AR (Assertion Roulette) smells. In tests with AR smells, the LLM successfully adds descriptive messages to assertions that lack them, which is attributed to its inherent code understanding capabilities. However, it does not eliminate all AR smells. When multiple smells are present in a test, using the default prompt (as shown in Figure 6) often leads to incomplete smell elimination, leaving AR smells unresolved. This issue is mitigated in UTREFACTOR by using a checkpoint mechanism, which

Table 5. Comparison results of test smell elimination effectiveness(#TS Num - number of Test Smells).

Project	#TS Num	UTREFACTOR		Llama3-70B		TESTAXE	
		aft.	rate	aft.	rate	aft.	rate
Commons-cli	149	21	↓ 86%	94	↓ 37%	-	-
Commons-compress	263	16	↓ 94%	55	↓ 79%	261	↓ <1%
Commons-math	196	28	↓ 86%	107	↓ 45%	158	↓ 19%
Gson	788	87	↓ 89%	375	↓ 52%	-	-
Jfreechart	324	44	↓ 86%	197	↓ 39%	-	-
Jsoup	655	69	↓ 91%	252	↓ 62%	635	↓ <1%
Total	2,375	69	↓ 89%	1,080	↓ 55%	2,315	↓ <1%

ensures that each smell is checked and eliminated after refactoring. For other types of smells, the default LLM's ability is also limited. For example, in handling exception-related smells, without the structured refactoring DSL rules designed in UTREFACTOR, the *Llama3-70B* often tends to simply replace throws statements with try-catch blocks, which does not effectively remove the smell. Overall, the default LLM exhibits more randomness in test refactoring. This characteristic becomes more prominent when multiple smells are present in a single test, resulting in unpredictable refactoring outcomes. In contrast, UTREFACTOR, with its integration of external knowledge and structured refactoring rules, consistently delivers more effective and stable smell elimination.

Effectiveness of Assertion Roulette Elimination. Adding any string message to assertions can eliminate the Assertion Roulette (AR) smell. However, the generated messages should clearly convey the intent of the assertion. To evaluate the quality of these messages, we conduct an experiment using GPT-4o, which is tasked with evaluating messages based on two criteria: **① Is the message clear and explicit?** **② Does the message align with the intent of the assertion?** In this experiment, we collect all 1,230 generated assertion messages. The results show that only 43 messages (3.5%) are flagged as **Unclear**, and all of these messages fail to meet criterion **①** (clarity of the message). After manual review, we find that these cases share a common feature: they lack sufficient context. For example, numerical variables (such as Magic Numbers) are extracted as constants with meaningful names. `assertEquals(expectedDatasetIndex, li.getDatasetIndex(), "The dataset index of the legend item should be 1")`, GPT-4o marks the message as unclear because it misinterprets `expectedDatasetIndex` as a variable rather than a constant. However, the generated message is reasonable, as the intent behind it is clear. Therefore, we conclude that LLMs can generate clear and descriptive assertion messages that align with the intent of the assertion.

Ablation of the DSL. We conduct an ablation study by comparing the effectiveness of UTREFACTOR with a version excluding the DSL, while keeping all other aspects the same (e.g., detailed descriptions of test smells, test context, and the smell checkpoint mechanism). The results show that without the DSL the reduction rate of test smells is 67%, which is an improvement over using LLM alone (55%) but significantly lower than the 89% achieved with the DSL.

Additionally, the natural language descriptions of test smells allow the LLM to partially address them effectively. For example, in 509 cases of Exception Catching Throwing, 175 cases (34%) successfully eliminate the smell using `assertDoesNotThrow` or `assertThrow`. However, the remaining cases show randomness, such as removing try-catch blocks or replacing throws with try-catch. In contrast, the DSL clearly defines rules for eliminating such smells, covering different scenarios (e.g., adding exception descriptions), leading to more consistent results.

Finally, we analyze all refactored tests that fail at runtime after successful compilation. Most failures occur when handling Duplicate Assertions, as our DSL refactors the original tests into JUnit

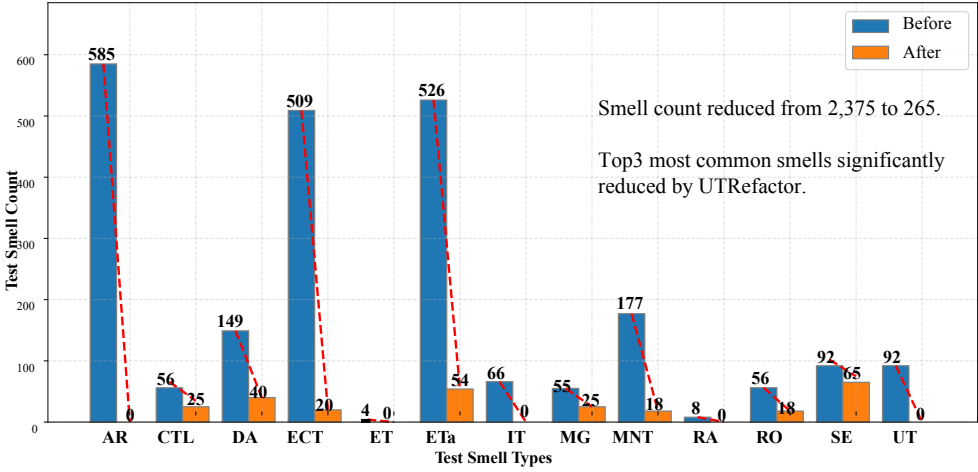


Fig. 8. Changes in the number of each type of test smell before and after test refactoring in UTREFACTOR (MG=Mystery Guest, RO=Resource Optimism, MNT=Magic Number Test, SE=Sensitive Equality, ETa=Eager Test, DA=Duplicate Assert, CTL=Conditional Test Logic, ECT=Exception Catching Throwing, AR=Assertion Roulette, RA=Redundant Assertion, IT=Ignored Test, UT=Unknown Test, ET=Empty Test).

5 parameterized tests. The LLM occasionally introduces errors in the parameterized data provided via @CsvSource, causing runtime failures. We find that the LLM struggles with converting tests into the parameterized format.

4.4 RQ3: How effective is UTREFACTOR in eliminating each category of Test Smell?

In this research question, we focus on UTREFACTOR’s ability to eliminate different types of test smells. Refactoring based on the type of code smells can lead to the same test being refactored multiple times, as many smells often coexist within a single test. Additionally, the order in which refactoring is performed can influence the final outcome (sec 3.2.3). To address these challenges, our refactoring process targets individual tests and simultaneously eliminates all smells. We first detect and collect all types of test smells present in the six open-source projects. We then analyze how effectively UTREFACTOR removes these smells.

As noted in RQ2, UTREFACTOR successfully reduces the total number of test smells from 2,375 to 265. Figure 8 provides detailed data on UTREFACTOR’s ability to eliminate each type of test smell. From the figure, we observe that UTREFACTOR completely eliminates the most prevalent type, i.e., AR (Assertion Roulette) smell. This success is due to UTREFACTOR’s integration of LLM code understanding capabilities, enabling it to infer appropriate descriptive messages based on the specific assert statements and surrounding code context. These messages help developers quickly identify the cause of errors when tests fail. Additionally, UTREFACTOR shows significant refactoring results for the two other high-frequency smells, ECT (Exception Catching Throwing) and ETa (Eager Test), reducing their numbers from 526 and 509 to 54 and 20, respectively.

We manually review the refactored tests generated by UTREFACTOR and analyze its effectiveness in eliminating each type of smell. Taking ETa (Eager Test) as an example, this smell occurs when a test includes assertions for multiple production functions, making the test’s purpose unclear and violating the best practice of the single responsibility principle. The refactoring approach for this smell is to split the test based on the production functions it verifies. UTREFACTOR’s advantage in eliminating ETa is demonstrated in two ways. First, its ETa DSL clearly defines how to identify API

assertions, split the test, and generate new test methods. Second, with the help of LLM, UTREFACTOR can understand and reason based on context, effectively handling complex code structures while avoiding the randomness typically associated with LLM-generated tests.

For the cases where smell elimination failed, we conducted a manual review and analysis. From the Figure 8, UTREFACTOR's ability to handle certain less frequent smells is somewhat limited. A notable example is SE (Sensitive Equality), which occurs when the default *toString* method is used in assertions. The refactoring approach for this smell requires either the project to have overridden the *toString* method or to provide an equivalent method for object comparison. In other words, if the project neither overrides *toString* nor offers a functionally equivalent method, UTREFACTOR is unable to remove this type of smell. Overall, across the six open-source projects, UTREFACTOR shows less effectiveness in addressing this smell type, primarily due to the lack of overridden *toString* methods or alternative methods in the project.

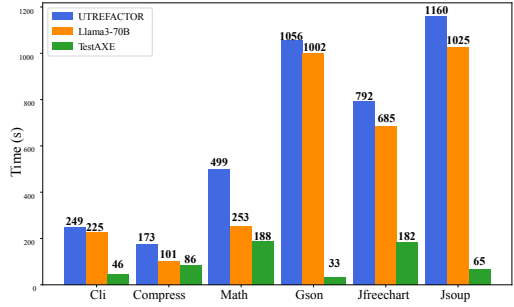


Fig. 9. Comparison of time costs between UTREFACTOR and two baseline tools across each project.

4.5 RQ4: What is the time efficiency of UTREFACTOR in test refactoring?

In this research question, we focus on the time efficiency of UTREFACTOR during the test refactoring process. We first collect the time data for each step of UTREFACTOR and then compare the total time with the two baseline tools.

Table 6. Time consuming(s) for each step of refactoring tests in UTREFACTOR.

Project	Step1	Step2	Step3
Cli	24	4	221
Compress	69	3	101
Math	258	6	235
Gson	64	4	988
Jfreechart	90	5	697
Jsoup	31	2	1,127
Total	536 (12%)	24 (<1%)	3,369 (87%)

Table 6 illustrates the time spent by UTREFACTOR at each refactoring step. Across all six open-source projects, the total time from detecting test smells to refactoring the 879 tests with identified smells is 4,379 seconds, averaging 3.8 seconds per test. This highlights UTREFACTOR's efficiency in terms of time consumption. Moreover, to accommodate various real-world refactoring scenarios, such as automatically refactoring test smells immediately after developers write a test or refactoring the entire test code of an existing project, UTREFACTOR supports refactoring at different levels: a single test, a single test file, or all test files within a project.

5 Discussion

5.1 Strengths of UTREFACTOR

In UTREFACTOR, we explore how LLMs perform automatic test refactoring tasks, guided by rules described in DSLs. This approach simulates the process that developers or testers follow during refactoring, from understanding the test's intent and identifying test smells to following step-by-step instructions outlined in the DSL. Compared to directly using an LLM for test refactoring, UTREFACTOR significantly reduces the risk of random errors, which can make the refactoring process unpredictable.

To further illustrate this issue, we randomly select 100 cases where the LLM fails to eliminate smells and conduct an analysis. Among them, 65 cases fail due to the LLM's arbitrary refactoring behavior, such as removing throws declarations or try-catch blocks without resolving the Exception Catching Throwing smell. 22 cases fail due to hallucinations during the elimination of the Eager Test smell, where the LLM introduces non-existent functions or variables into split tests. 23 cases fail because the LLM generates helper methods with empty bodies and comments instructing users to complete them.

5.2 Integration of UTREFACTOR with LLMs

Closed-source LLMs often come with high token usage costs, and their large parameter sizes lead to significant deployment expenses. To address this, UTREFACTOR uses the open-source and free *Llama3-70B* model. However, considering the tool's extensibility, UTREFACTOR also supports replacing *Llama3-70B* with other LLM models. *ChatGPT* [2] is a widely recognized model and the current stable version *GPT-4o* [7]. We design an experiment to demonstrate that UTREFACTOR effectively eliminates test smells using alternative LLM models. As shown in Table 7, we compare two setups: one using the default *GPT-4o* with the general prompt (as illustrated in Figure 7) and another using UTREFACTOR with *Llama3-70B* replaced by *GPT-4o*. We randomly select 100 tests with smells from six open-source datasets, containing a total of 251 test smells. After refactoring the tests, Table 7 shows that using *GPT-4o* with the default prompt achieves a smell reduction rate of 55%, whereas UTREFACTOR achieves a reduction rate of 91%.

Table 7. Effectiveness of UTREFACTOR integrated with *GPT-4o* in eliminating test smells.

Tool	#TS Num	aft.(rate)
Default <i>GPT-4o</i>	251	108 (↓57%)
UTREFACTOR (<i>GPT-4o</i>)	251	23 (↓91%)

We manually review the test refactoring results using *GPT-4o*. Without guidance from external knowledge on test smells, the default *GPT-4o* encounters issues in eliminating smells. For example, in the `testFindRangeBounds` of the *jfreechart*, where four types of smells are present. *GPT-4o* removes only three smells but fails to eliminate the most common smell, *Assertion Roulette*. In other cases, we observe

that *GPT-4o* attempts to address *Assertion Roulette* by adding comments to assertion statements instead of explicitly adding a message, which does not properly resolve the smell.

In contrast, UTREFACTOR leverages its DSL to clearly instruct the addition of a message to assertions and ensure that all smells are addressed during refactoring, avoiding the issues caused by the default LLM approach. In another example, the `testDrawWithNullInfo` contains a complex try/catch structure detected as *Exception Catching Throwing*. *GPT-4o* simply removes this block and adds an explicit throw in the method signature, but this fails to eliminate the smell. By contrast, guided by UTREFACTOR's DSL, the tool uses `assertNotThrow` to replace the try/catch block and successfully resolves the smell. This demonstrates that UTREFACTOR effectively eliminates smells even when using other LLMs like *GPT-4o*.

5.3 Threats to Validity

Threats to Internal Validity. To minimize hallucinations and reduce errors during the test smell elimination process, we provide LLMs with sufficient context related to test refactoring and design a refactoring DSL to guide the LLM step by step in removing test smells. However, the DSL in UTREFACTOR may not cover every possible scenario, which can impact the overall effectiveness of the test refactoring. Our refactoring DSL is designed as a flexible, external configuration file, making it easy to extend. For example, new smell types can be addressed by simply adding or updating DSL files, modifying refactoring steps to accommodate a broader range of scenarios.

Threats to External Validity. We integrate the *tsDetect* tool into our process for test smell detection. This tool uses a built-in Java parser to analyze test code and identify smells, currently supporting Java versions up to 13. For incompatible Java versions, *tsDetect* may encounter syntax parsing errors during detection, which could prevent the identification of test smells, thus impacting the effectiveness of smell elimination. Given that test smell detection tools are continually evolving, UTREFACTOR is designed to be highly extensible, allowing for the replacement of the built-in test smell detection tool. When encountering incompatible Java versions, a more robust detection tool can be integrated to address undetected smells, thereby improving the overall performance of the test refactoring process.

6 Related Work

The Impact and Refactoring of Test Smells. Several studies [20, 21, 26, 27, 29, 32, 33, 35–37] have focused on the impact of test smells, including how they affect the development process, software maintenance, and comprehension. To investigate how much developers acknowledge the presence of test smells, Soares et al. [36] conducted a study with 73 experienced open-source developers across 272 projects. They analyzed preferences and motivations related to 10 identified test smells by comparing the original and refactored versions of test code. The results showed that 78% of developers acknowledged the negative impact of test smells and preferred refactored tests. Besides, they explored the use of new JUnit 5 features to eliminate and prevent test smells in another study [37]. In an empirical study on 485 popular Java open-source projects from GitHub, Soares et al. found that only 15.9% of projects used the JUnit 5 library. By applying seven JUnit 5 features to address test smells, they conducted a survey of 212 developers and submitted 38 pull requests, achieving a 94% acceptance rate among respondents.

Test Smell Detection. There are many research methods and tools [18, 19, 22, 25, 30, 31, 34, 38–40] available for detecting test smells across different programming languages. Palomba et al. [30] developed *TASTE*, an automated textual-based tool for detecting several types of test smells. Compared to previous structure-based detection methods, this tool improved smell detection effectiveness by 44%. Peruma et al. [31] recently developed a tool called *tsDetect*, capable of detecting 19 test smells in Java. It used a set of detection rules to locate existing test smells in test code. Wang et al. [39] proposed a tool called *PYNOSE*, designed to detect 17 types of test smells within Python’s standard *unittest* framework, and introduced a new test smell type called *Suboptimal Assert*. *PYNOSE* was available as a plugin for *PyCharm*, and in an empirical study conducted across 248 Python projects, they found that 98% of projects contained at least one type of test smell. Unlike work focused on test smell detection, our research aims to explore how the latest large language models could improve the performance of eliminating detected smells.

7 Conclusion and Future Work

This paper introduces UTREFACTOR, a tool that focuses on the automatic detection and refactoring of test smells. In this work, we explore how leveraging refactoring DSLs and incorporating test context knowledge optimize and enhance the ability of open-source LLMs to automatically refactor test code. UTREFACTOR supports multiple test smell types and eliminates smells at varying granularities. Compared to existing methods, UTREFACTOR demonstrates greater effectiveness in eliminating various types of test smells from code. In the future, we plan to extend our refactoring rules to support additional programming languages, such as Python and C/C++. Additionally, we intend to investigate the impact of using different LLM sizes and parameters on the test refactoring process.

8 Data Availability

Our tool is available at [16].

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2024YFB4506400) and sponsored by CCF-Huawei Populus Grove Fund.

References

- [1] [n. d.]. Apache. <https://www.apache.org/>.
- [2] [n. d.]. ChatGPT. <https://chat.openai.com/>.
- [3] [n. d.]. Commons-cli. <https://github.com/apache/commons-cli>.
- [4] [n. d.]. Commons-compress. <https://github.com/apache/commons-compress>.
- [5] [n. d.]. Commons-math-legacy. <https://github.com/apache/commons-math/tree/master/commons-math-legacy>.
- [6] [n. d.]. CoT. <https://arxiv.org/abs/2201.11903>.
- [7] [n. d.]. GPT-4o. <https://openai.com/index/hello-gpt-4o/>.
- [8] [n. d.]. Gson. <https://github.com/google/gson>.
- [9] [n. d.]. Jacoco. <https://www.jacoco.org/jacoco/trunk/index.html>.
- [10] [n. d.]. Jfreechart. <https://github.com/jfree/jfreechart>.
- [11] [n. d.]. Jsoup. <https://github.com/jhy/jsoup>.
- [12] [n. d.]. Llama3-70B. <https://huggingface.co/meta-llama/Meta-Llama-3-70B>.
- [13] [n. d.]. Rascal. <https://www.rascal-mpl.org/>.
- [14] [n. d.]. SRP. https://en.wikipedia.org/wiki/Single-responsibility_principle.
- [15] [n. d.]. tsDetect. <https://github.com/TestSmells/TSDetect>.
- [16] [n. d.]. UTRefactor. <https://github.com/testmigrator/testsmellrefactoring>.
- [17] Mahmoud Alfadel, Diego Elias Costa, and Emad Shihab. 2023. Empirical analysis of security vulnerabilities in python packages. *Empirical Software Engineering* 28, 3 (2023), 59. <https://doi.org/10.1007/S10664-022-10278-4>
- [18] Wajdi Aljedaani, Anthony Peruma, Ahmed Aljohani, Mazen Alotaibi, Mohamed Wiem Mkaouer, Ali Ouni, Christian D Newman, Abdullatif Ghallab, and Stephanie Ludi. 2021. Test smell detection tools: A systematic mapping study. In *Proceedings of the 25th International Conference on Evaluation and Assessment in Software Engineering*. 170–180. <https://doi.org/10.1145/3463274.3463335>
- [19] Alexandru Bodea. 2022. Pytest-Smell: a smell detection tool for Python unit tests. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*. 793–796. <https://doi.org/10.1145/3533767.3543290>
- [20] Humberto Damasceno, Carla Bezerra, Denivan Campos, Ivan Machado, and Emanuel Coutinho. 2023. Test smell refactoring revisited: What can internal quality attributes and developers' experience tell us? *Journal of Software Engineering Research and Development* (2023), 13–1. <https://doi.org/10.5753/JSERD.2023.3195>
- [21] Humberto Damasceno, Carla Bezerra, Emanuel Coutinho, and Ivan Machado. 2022. Analyzing test smells refactoring from a developers perspective. In *Proceedings of the XXI Brazilian Symposium on Software Quality*. 1–10. <https://doi.org/10.1145/3571473.3571487>
- [22] Phongphan Danphitsanuphan and Thanitta Suwantada. 2012. Code smell detecting tool and code smell-structure bug relationship. In *2012 Spring congress on engineering and technology*. IEEE, 1–5.
- [23] Estevan Alexander de Paula and Rodrigo Bonifácio. 2022. TestAXE: Automatically Refactoring Test Smells Using JUnit 5 Features. In *Congresso Brasileiro de Software: Teoria e Prática (CBSOFT)*. SBC, 89–98.
- [24] Xueying Du, Geng Zheng, Kaixin Wang, Jiayi Feng, Wentai Deng, Mingwei Liu, Bihuan Chen, Xin Peng, Tao Ma, and Yiling Lou. 2024. Vul-RAG: Enhancing LLM-based Vulnerability Detection via Knowledge-level RAG. *arXiv preprint arXiv:2406.11147* (2024). <https://doi.org/10.48550/ARXIV.2406.11147>
- [25] Daniel Fernandes, Ivan Machado, and Rita Maciel. 2022. TEMPY: Test smell detector for Python. In *Proceedings of the XXXVI Brazilian Symposium on Software Engineering*. 214–219. <https://doi.org/10.1145/3555228.3555280>
- [26] 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 *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 171–182. <https://doi.org/10.1109/ICSME52107.2021.00022>
- [27] Luana Martins, Taher A Ghaleb, Heitor Costa, and Ivan Machado. 2024. A comprehensive catalog of refactoring strategies to handle test smells in Java-based systems. *Software Quality Journal* (2024), 1–39. <https://doi.org/10.1007/S11219-024-09663-7>
- [28] Marjan Mernik, Jan Heering, and Anthony M Sloane. 2005. When and how to develop domain-specific languages. *ACM computing surveys (CSUR)*, 37, 4 (2005), 316–344. <https://doi.org/10.1145/1118890.1118892>
- [29] Nicholas Alexandre Nagy and Rabe Abdalkareem. 2022. On the co-occurrence of refactoring of test and source code. In *Proceedings of the 19th International Conference on Mining Software Repositories*. 122–126. <https://doi.org/10.1145/3524842.3528529>

- [30] Fabio Palomba, Andy Zaidman, and Andrea De Lucia. 2018. Automatic test smell detection using information retrieval techniques. In *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 311–322. <https://doi.org/10.1109/ICSME.2018.00040>
- [31] Anthony Peruma, Khalid Almalki, Christian D Newman, Mohamed Wiem Mkaouer, Ali Ouni, and Fabio Palomba. 2020. Tsdetect: An open source test smells detection tool. In *Proceedings of the 28th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*. 1650–1654.
- [32] Adriano Pizzini. 2022. Behavior-based test smells refactoring: Toward an automatic approach to refactoring eager test and lazy test smells. In *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings*. 261–263. <https://doi.org/10.1145/3510454.3517059>
- [33] Adriano Pizzini, Sheila Reinehr, and Andreia Malucelli. 2023. Sentinel: A process for automatic removing of Test Smells. In *Proceedings of the XXII Brazilian Symposium on Software Quality*. 80–89.
- [34] Valeria Pontillo, Dario Amoroso d’Aragona, Fabiano Pecorelli, Dario Di Nucci, Filomena Ferrucci, and Fabio Palomba. 2024. Machine learning-based test smell detection. *Empirical Software Engineering* 29, 2 (2024), 55. <https://doi.org/10.48550/ARXIV.2208.07574>
- [35] Railana Santana, Luana Martins, Tássio Virgínio, Larissa Rocha, Heitor Costa, and Ivan Machado. 2024. An empirical evaluation of RAIDE: A semi-automated approach for test smells detection and refactoring. *Science of Computer Programming* 231 (2024), 103013. <https://doi.org/10.1016/J.SCICO.2023.103013>
- [36] Elvys Soares, Márcio Ribeiro, Guilherme Amaral, Rohit Gheyi, Leo Fernandes, Alessandro Garcia, Balduino Fonseca, and André Santos. 2020. Refactoring test smells: A perspective from open-source developers. In *Proceedings of the 5th Brazilian Symposium on Systematic and Automated Software Testing*. 50–59.
- [37] Elvys Soares, Marcio Ribeiro, Rohit Gheyi, Guilherme Amaral, and André Santos. 2022. Refactoring test smells with junit 5: Why should developers keep up-to-date? *IEEE Transactions on Software Engineering* 49, 3 (2022), 1152–1170. <https://doi.org/10.1109/TSE.2022.3172654>
- [38] Tássio Virgínio, Luana Martins, Larissa Rocha, Railana Santana, Adriana Cruz, Heitor Costa, and Ivan Machado. 2020. Jnose: Java test smell detector. In *Proceedings of the XXXIV Brazilian Symposium on Software Engineering*. 564–569. <https://doi.org/10.1145/3422392.3422499>
- [39] Tongjie Wang, Yaroslav Golubev, Oleg Smirnov, Jiawei Li, Timofey Bryksin, and Iftekhar Ahmed. 2021. Pynose: a test smell detector for python. In *2021 36th IEEE/ACM international conference on automated software engineering (ASE)*. IEEE, 593–605. <https://doi.org/10.1109/ASE51524.2021.9678615>
- [40] Yanming Yang, Xing Hu, Xin Xia, and Xiaohu Yang. 2024. The Lost World: Characterizing and Detecting Undiscovered Test Smells. *ACM Transactions on Software Engineering and Methodology* 33, 3 (2024), 1–32. <https://doi.org/10.1145/3631973>
- [41] Zhiqiang Yuan, Mingwei Liu, Shiji Ding, Kaixin Wang, Yixuan Chen, Xin Peng, and Yiling Lou. 2024. Evaluating and improving chatgpt for unit test generation. *Proceedings of the ACM on Software Engineering* 1, FSE (2024), 1703–1726. <https://doi.org/10.1145/3660783>

Received 2024-09-13; accepted 2025-01-14