

Enhancing IR-based Fault Localization using Large Language Models

Shuai Shao
University of Connecticut
Storrs, CT, USA

Tingting Yu
University of Connecticut
Storrs, CT, USA

ABSTRACT

Information Retrieval-based Fault Localization (IRFL) techniques aim to identify source files containing the root causes of reported failures. While existing techniques excel in ranking source files, challenges persist in bug report analysis and query construction, leading to potential information loss. Leveraging large language models like GPT-4, this paper enhances IRFL by categorizing bug reports based on programming entities, stack traces, and natural language text. Tailored query strategies, the initial step in our approach (LLMiRQ), are applied to each category. To address inaccuracies in queries, we introduce a user and conversational-based query reformulation approach, termed LLMiRQ+. Additionally, to further enhance query utilization, we implement a learning-to-rank model that leverages key features such as class name match score and call graph score. This approach significantly improves the relevance and accuracy of queries. Evaluation on 46 projects with 6,340 bug reports yields an MRR of 0.6770 and MAP of 0.5118, surpassing seven state-of-the-art IRFL techniques, showcasing superior performance.

ACM Reference Format:

Shuai Shao and Tingting Yu. 2024. Enhancing IR-based Fault Localization using Large Language Models. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/nnnnnnnnnnnnnnn>

1 INTRODUCTION

Fault localization plays a crucial role in software maintenance and development processes. When a bug is discovered by a user or a software developer, it is typically reported in a bug tracking system, such as Bugzilla [1], Google Code Issue Tracker [3], or Github Issue Tracker [2], through a bug report or an issue report. These reports offer valuable information that assists developers in efficiently identifying and resolving bugs.

Upon receiving a bug report, a developer's primary tasks include reproducing and localizing the bug within the source code. If we consider the bug report as a query and the source code files in the software repository as a collection of documents, the challenge of identifying relevant source files for a given bug report aligns with a standard task in information retrieval (IR). Several approaches have been proposed in IR-based fault localization [28, 33, 35, 41], where

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 ACM 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.

Conference'17, July 2017, Washington, DC, USA
© 2024 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00
<https://doi.org/10.1145/nnnnnnnnnnnnnnn>

similarity scores between bug reports and program entities (e.g., source files) are computed. Subsequently, rankings derived from these similarity scores reveal the relevance of software entities to the specific bug report.

While IR-based fault localization (IRFL) has undergone extensive research and development, it encounters significant challenges. A recent study by Lee et al. [18] assessed the reproducibility of IRFL performance. Despite employing various retrieval methods, resources, and query formulation strategies in six investigated techniques—BugLocator, BLUiR, BRTtracer, AmaLgam, BLIA, and Locus—no significant performance differences were observed.

A key challenge in IRFL lies in constructing effective queries for retrieval. Common approaches view bug reports as queries, often involving tokenization and stop word removal [26, 28, 35, 39]. However, these methods frequently retain noise within the query. Additionally, existing query formulation strategies assign different weights to bug report components, such as the title, description, and steps to reproduce, for query construction [6, 7]. Some strategies leverage graph creation and term weighting to identify essential tokens for query construction [25, 26]. Despite these efforts, these methods may lack a holistic understanding of the semantic content in bug reports, leading to noise introduction or critical information oversight.

The emergence of Large Language Models (LLMs) [5, 31] like ChatGPT has revolutionized natural language understanding. LLMs excel in interpreting natural language nuances and contexts, offering a promising avenue for deeper bug report analysis and construction of effective queries to enhance IRFL performance.

This paper introduces LLMiRQ, an innovative approach for automated fault localization utilizing a learning-to-rank machine learning framework [16]. Learning-to-rank has been adopted by existing work for both IR-based [38] and spectrum-based [4] fault localization. The key idea of using Learning-to-Rank for fault localization is to train a machine learning model to rank pieces of software code by how likely they are to contain a bug, based on their relevance to a given bug report. For example, Ye et al. [38] employed a Learning-to-Rank approach to prioritize relevant files in bug reports through the integration of domain knowledge.

LLMiRQ defines its ranking function as a weighted combination of features that comprehensively capture bug report semantic meanings, distinguishing it from existing techniques. Specifically, LLMiRQ harnesses the state-of-the-art Large Language Model (LLM), GPT-4, for this purpose. Moreover, LLMiRQ proposes new features tailored for learning-to-rank models based on the newly constructed queries by LLMs. Additionally, we present LLMiRQ+ as an enhanced version integrating GPT-4's conversational capabilities for iterative query reformulation based on user feedback, enhancing query accuracy. Evaluation against seven contemporary IRFL techniques using a 6,340-bug report dataset demonstrates

LLmiRQ+'s superior performance. With a high Mean Reciprocal Rank (MRR) of 0.6770 and Mean Average Precision (MAP) of 0.5118, LLmiRQ+ outperforms all seven state-of-the-art IRFL techniques. Notably, LLmiRQ pioneers LLM application to query construction in IRFL.

In summary, this paper makes the following contributions:

- A novel query construction approach for IRFL using Large Language Models (LLMs).
- Development of a user feedback-driven query reformulation process, including a bias-mitigation template, improving query precision and reliability.
- Strategic use of key features and a learning-to-rank model to enhance IRFL, ensuring accurate fault localization.
- Provision of a tool and experimental dataset publicly available¹.

2 BACKGROUND AND MOTIVATION

In this section, we describe the background and provide motivating examples.

2.1 Background

Information Retrieval (IR)-based fault localization is a technique that utilizes information retrieval methods to pinpoint potential bug locations in software. This approach capitalizes on the textual information surrounding software elements, such as source code, comments, and error logs, as it often contains crucial insights into the functionality and potential faults of these elements. By treating source code and bug reports as textual documents, IR-based fault localization employs text similarity measures to align the descriptions in bug reports with the corresponding parts of the source code.

The pivotal components of IRFL encompass the query and retrieval models, where significant research has been dedicated to enhancing and refining both aspects. For improving query construction, Saha et al.[28] utilized the Eclipse JDT to effectively process the source code's Abstract Syntax Tree (AST), extracting vital information from various code entities. This significantly enhances the efficiency of IR-based fault localization by constructing more precise queries. Rahman et al. [26] incorporates context-aware query reformulation into bug localization, using a graph-based approach to represent bug reports and employing PageRank to identify important tokens for queries, demonstrating some improvements. However, as highlighted in subsequent sections, existing queries often *lack the accuracy or comprehensiveness* necessary for effective fault localization.

Efforts have also been made to enhance retrieval models in fault localization [26, 28, 33, 39]. For example, Wang et al.[33] proposed a method that leverages version history information to refine IRFL techniques, thereby improving the model's ability to adeptly exploit historical data. Beyond static retrieval models, learning-based models are also utilized, such as Learning-to-Rank (Ltr) [16]. Ltr represents a machine learning methodology used to develop ranking models for information retrieval systems. The Ltr process begins with feature extraction from bug reports and source code, including textual similarity scores and code metrics. By leveraging

¹<https://github.com/jedishao/LLmiRQ>

Table 1: Motivation Examples of Bug Reports and Queries.

COMPRESS-357
BugLocator: title + description Rank: 11
BLIZZARD: finish stream calls output method class synchronize time bzip compressor compressoroutputstream output stream compressor stream affect incorrectly Rank: 47
GPT-4: BZip2CompressorOutputStream finalize finish Rank: 1
CAMEL-620
Title: ResequencerType.createProcessor could throw NPE as stream config does not get initialized.
Description: java.lang.NullPointerException at org.apache.camel.model.ResequencerType.createProcessor(ResequencerType.java:163) at org.apache.camel.model.ProcessorType.createOutputsProcessor(ProcessorType.java:584) at org.apache.camel.model.ProcessorType.createOutputsProcessor(ProcessorType.java:93) ...
BugLocator: title + description Rank: 223
BLIZZARD: NullPointerException createStreamResequencer createProcessor create createOutputsProcessor ResequencerType addRoutes ProcessorType Processor RouteType startRouteDefinitions InterceptorRef Rank: 2659
GPT-4: ResequencerType ResequencerTest createProcessor Rank: 2
CAMEL-2320
Title: JDBC component doesn't preserve headers
Description: JDBC component doesn't preserve any of the headers that are sent into it
BugLocator: title + description Rank: 179
BLIZZARD: set jdbc ftp row parameters generated endpoint output JDBC component preserve headers JDBC component preserve headers Rank: 7
GPT-4: JdbcProducer JdbcEndpoint JdbcComponent Rank: 3

historical bug reports and their corresponding fixes, training data is constructed. This data comprises queries (bug reports), candidate documents (source code files or methods), and relevance labels indicating the presence of faults. This data is used to train the Ltr model to rank candidate documents based on their likelihood of containing faults.

Several methodologies have surfaced that harness Learning-to-Rank (Ltr) for fault localization from bug reports. For instance, Ye et al. [38] utilized a Learning-to-Rank strategy to prioritize pertinent files in bug reports by integrating domain knowledge. Likewise, Le et al. [4] applied Learning-to-Rank techniques to bolster spectrum-based fault localization methodologies. Despite these advancements, current approaches grapple with suboptimal accuracy primarily due to *imprecise feature representation* that fails to accurately capture the mapping between bug reports and source code.

2.2 Motivation

2.2.1 Query Construction. Table 1 demonstrates the limitations of current methods and the potential improvements using GPT-4 for different types of bug reports.

In COMPRESS-357, traditional methods like BugLocator and BLIZZARD struggle with noisy data. GPT-4, however, understands the bug report and pinpoints the root cause by identifying crucial class names like 'BZip2CompressorOutputStream', leading to more accurate rankings. For CAMEL-620, which includes stack traces, existing tools often fail to analyze them effectively. This is because they primarily focus on the position, such as the top 10 entries of stack traces, without truly understanding the context or the content of the stack trace itself. GPT-4 can analyze the stack trace and identify the root cause by pointing to specific classes and methods, enhancing fault localization accuracy. In CAMEL-2320, the bug report contains only natural language with little useful information. Traditional methods fall short here, but GPT-4 can extend the information by understanding the context and identifying

relevant classes and methods, thereby improving the accuracy of fault localization.

2.2.2 Retrieval Model. Furthermore, for bug reports CAMEL-620 and CAMEL-2320 in Table 1, the root causes pinpointed by GPT-4—ResequencerType and JdbcProducer, respectively—do not achieve the top rank position when using conventional retrieval models like BugLocator’s rVSM. The challenge lies in the nature of queries generated by GPT-4, which consist of programming entities. Traditional retrieval models that calculate similarity between these queries and the source code often fail to filter out noise present in the source code, such as comments or irrelevant programming entities. This observation leads to the hypothesis that a *more precise* approach, focusing on exact matches of class and method names, could yield improvements.

Additionally, empirical investigations suggest a tendency for files implicated in a single bug report to exhibit call relationships [19]. For instance, in the case of CAMEL-620, not only is ResequencerType identified as the buggy file, but associated files like ResequencerTest and SteamResequencerType, which interact with ResequencerType, also require attention. This interconnection underlines the potential benefits of considering call relationships in enhancing retrieval models.

In summary, traditional retrieval models are not optimally aligned with the unique characteristics of queries generated by advanced language models like GPT-4. To bridge this gap, we propose an *enriched feature set* for a learning-to-rank model, encompassing class name match score, call graph score, and five features adopted from existing work [38]. Integrating these features into a learning-to-rank framework is expected to significantly augment the efficacy of IR-based fault localization.

3 APPROACH

Figure 1 provides an overview of our approach. Given that different bug report components, such as programming entities and stack traces, can have significantly varying impacts on IRFL, tailored approaches are required for analysis. To begin, we segment bug reports into three distinct categories based on their content. We first determine if a report contains a stack trace; if it does, we categorize it as ST (Stack Trace). If not, we then check for the presence of programming entities; if present, the report is categorized as PE (Programming Entities). Reports that contain neither stack traces nor programming entities are classified as NL (Natural Language).

The rationale behind the categories is that for PE and ST bug reports, which often contain precise details such as specific class or method names, a *query reduction* strategy is applied to minimize noise and highlight the critical information. In contrast, NL reports usually lack direct code references, necessitating a *query extension* approach to incorporate broader contextual information for effective fault localization. This approach is referred to as LLmiRQ.

Each category is subjected to a customized query construction strategy, which serves as the initial query for IRFL. Our approach uses iterative design of prompts for each category to optimize the query construction process: Prompts are initially designed based on the type of information present in the bug report. Prompts are evaluated and refined based on feedback and empirical results to improve their effectiveness. The refined prompts are used to

generate precise queries tailored to the content of each bug report category. The detail of the prompts are discussed in Section 3.2.

Upon generating the initial IRFL query, we obtain a results list. In cases where the initial IRFL results fail to precisely identify the problematic files, we introduce a *user and conversation-driven query reformulation* process facilitated by GPT. This enhanced approach is denoted as LLmiRQ+. Through this iterative refinement, we aim to achieve more accurate and effective fault localization.

Besides query construction, our approach enhances IR-based fault localization by leveraging a learning-to-rank (Ltr) model, which is trained using historical bug reports and their corresponding fixes. The Ltr model ranks the source files, providing users with a list where higher-ranked files are more likely to contain the bug. We design two novel features to capture various aspects of the bug reports and source code: the *class name match score* and the *call graph score*. The class name match score helps identify relevant classes involved in the bug, while the call graph score captures dynamic relationships between code components. Additionally, we incorporate five features from existing Ltr works, such as text similarity, historical bug fix frequency, and recency, to enhance our model’s effectiveness.

3.1 Bug Reports Classification

As described in Section 2, the quality of bug reports plays an important role in IR-based fault localization. Different kinds of bug reports need different query construction strategies.

3.1.1 Text Containing Programming Entities (PE). Program elements include method names, package names, and source file names. Bug reports containing one or more of these elements, but no stack traces, are classified as **PE**. Queries from these reports are rich in information. Appropriate regular expressions [27] are applied to identify these elements within the text.

3.1.2 Stack Traces (ST). Stack traces include sequences of method calls during an error, often pinpointing the exact class or method where the error occurred. Bug reports with one or more stack traces are labeled as **ST**. Since stack traces contain structured information, queries generated can be noisy. We use specific regular expressions [21] to identify relevant trace entries.

3.1.3 Pure Text (NL). Bug reports described solely in plain natural language, without specific references to programming entities or stack traces, are classified as **NL**. These reports pose challenges due to the lack of direct code-related terms, requiring reliance on the textual content’s accuracy.

We build a classifier by leveraging previous work [26] to automatically categorize bug reports. Using this classifier, LLmiRQ flags a given bug report into a specific category and applies the corresponding strategy to construct the query (Figure 1).

3.2 Query Construction

By leveraging GPT’s ability to comprehend and contextualize diverse textual information, our goal is to explore its potential in assisting developers in efficiently identifying bugs. This section outlines our three core query construction strategies: query reduction, query expansion, and query reformulation, and describes their utilization in IR-based fault localization.

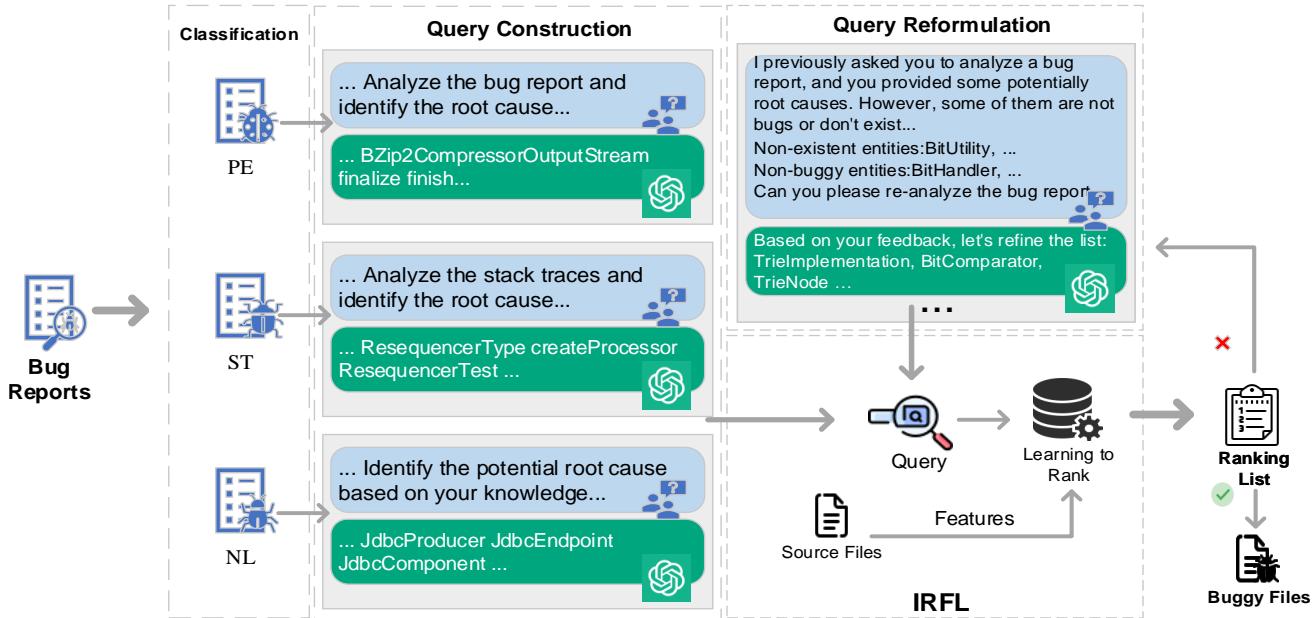


Figure 1: The overview of LLmiRQ and LLmiRQ+.

3.2.1 Query Reduction. Query reduction aims to highlight the most important tokens and discard superfluous ones, a task where traditional methods often fall short. GPT-4's advanced contextual understanding enhances the precision of identifying root causes, thus revolutionizing query reduction. The prompt used is:

"Analyze the bug report and construct a query by identifying programming entities (e.g., classes, methods) that may be relevant to the bug's root cause."

This prompt was iteratively designed through the following steps: *Initial Query Construction*: The initial prompt focused broadly on identifying key elements in the bug report. Early versions did not specify the type of programming entities, which resulted in less targeted queries. *Feedback and Refinement*: Feedback indicated the need for more precise identification of programming entities. As a result, the prompt was refined to specify "classes" and "methods" explicitly, directing the model's attention to these crucial elements. *Empirical Testing*: The refined prompt was tested on various bug reports to ensure it consistently produced accurate and relevant queries. For example, in bug report COMPRESS-357, GPT-4 generates the query: "BZip2CompressorOutputStream finalize finish," derived from the analysis of the bug report.

For stack traces (ST), which inherently carry crucial information, GPT-4 can accurately identify relevant classes and methods. The prompt used is:

"Analyze the provided stack traces and construct a query, identifying programming entities (e.g., classes, methods) relevant to the bug's root cause."

This prompt was iteratively designed as follows: *Initial Query Construction*: The initial prompt was general and did not focus specifically on stack traces. It was found to be less effective in extracting relevant information from stack traces. *Feedback and Refinement*: The prompt was adjusted to explicitly mention "stack

traces" and to identify "classes" and "methods," which are often the key elements in stack traces. *Empirical Testing*: The refined prompt was tested on bug reports with stack traces, demonstrating improved accuracy in identifying relevant programming entities.

In bug report CAMEL-620, the query generated includes entities like "ResequencerType ResequencerTest createProcessor."

3.2.2 Query Expansion. Query expansion is essential for bug reports that lack informative details (NL), such as specific programming entities or stack traces. Traditional methods require significant human effort and often do not consistently improve IRFL outcomes. GPT-4, with its vast internal knowledge, can introduce relevant programming entities, enriching the available information. The prompt used is:

"Analyze the bug report and construct a query by identifying potential programming entities (e.g., classes, methods) relevant to the bug's root cause based on your knowledge."

This prompt was iteratively designed through the following steps: *Initial Query Construction*: The initial prompt was very broad and did not effectively guide the model to identify potential programming entities. *Feedback and Refinement*: Feedback highlighted the need for the model to draw upon its extensive internal knowledge. The prompt was refined to specify "potential programming entities (e.g., classes, methods)" to guide the model more effectively. *Empirical Testing*: The refined prompt was tested on bug reports lacking specific details, showing improved performance in generating insightful queries.

After expanding the query, a straightforward codebase analysis is conducted to ensure the suggested classes or methods exist within the project. Non-existent entities are removed. For example, in bug report CAMEL-2320, GPT-4 might suggest entities like "JdbcProducer," "JdbcEndpoint," and "JdbcComponent," inferred from the context of the bug report.

By adopting these tailored strategies and iteratively designing prompts, our approach enhances the overall performance of IRFL, addressing the limitations of traditional models and making full use of the precise queries generated by GPT-4.

3.3 Interactive Query Reformulation

When the initial query does not produce satisfactory results, reformulating the query could help optimize the search process. Inspired by existing works [7, 13, 37], we have designed a user and conversational-driven query reformulation using GPT. The envisioned usage scenario begins with the execution of an initial query. Subsequently, users examine the results list (the top 10 files) and, if the buggy files are not present in this list, they can provide feedback to GPT.

To minimize bias in the recommendations provided by LLMs while maintaining the feedback's informativeness and value, we've designed two types of feedback mechanisms. The first is **Feedback on Non-Existing Classes**, where users report classes suggested by the AI that don't exist in the source code. For example, consider a situation like the query reformulation depicted in Figure 1. In this scenario, a user provides feedback stating, "I couldn't find any class named BitUtility in the source code." Such feedback assists the model in refining its suggestions by excluding non-existing or irrelevant classes, thereby improving the precision of its recommendations. The second type of feedback is **Feedback on Non-Buggy Classes**. In this case, users inform the model about classes that were suggested but are not related to the bug (non-buggy). For instance, refer to the second feedback provided in Figure 1, where the user states, "TrieImplementation doesn't seem to have the issue." The model utilizes this feedback to narrow down its suggestions, concentrating on classes that are more likely to be buggy. This approach enhances the relevance and utility of the model's recommendations.

Taking into account both user experience and computational efficiency, we've established a maximum of 5 conversation iterations. Persistently soliciting feedback without yielding satisfactory results can prove to be both time-consuming and exasperating for users. If the buggy files remain elusive even after 5 query reformulations based on user insights, it may suggest inherent shortcomings in the query formulation methodology or inconsistencies in the underlying data. In such cases, the query reformulation is deemed unsuccessful. Conversely, if a reformulated query successfully pinpoints the buggy files within the result list, then the query reformulation is considered a success.

3.4 Feature Selection

Common Information Retrieval-based Fault Localization (IRFL) techniques often utilize bug reports as queries. These reports typically include extraneous information that introduces noise into the retrieval process. In contrast, our method constructs queries composed exclusively of programming entities. This focused approach necessitates a departure from traditional text similarity measures, which may also amplify noise due to their inclusion of less relevant textual data in source code. To harness the full potential of our refined queries, we have carefully curated a set of features designed to capitalize on the precise nature of the programming entities

identified by GPT-4. These features aim to enhance the alignment between query terms and relevant source code, thereby improving the accuracy and efficiency of the fault localization process.

3.4.1 Class Name Match Score. Given that our queries are comprised solely of programming entities, specifically class and method names, these entities offer a direct and potent indication of potentially buggy files. Our first selected feature, the Class Name Match Score, is designed to exploit this correlation. For each source file s that contains a class name appearing in our query q , we calculate this feature based on the length of the class name in question ($s.class$).

We opt for a score proportional to the name's length, rather than a binary presence indicator (1 or 0), to facilitate a more nuanced ranking in the fault localization process. This decision stems from the premise that longer class names tend to be more specific. Consequently, a bug report that explicitly mentions a lengthy and specific class name is likely pointing directly towards the buggy file. Therefore, by evaluating files based on the length of the matched class name, we aim to prioritize files with more specific (and thus potentially more relevant) class names in the ranking list.

$$f_1(s, q) = |s.class| \quad \text{if } s.class \in q$$

This approach ensures that the ranking not only highlights files with matching class names but also subtly distinguishes them based on the specificity implied by the length of these names, thereby enhancing the effectiveness of the fault localization.

3.4.2 Call Graph Score. The textual similarity between source files, though insightful, may not fully capture the complexity of their interdependencies. Our empirical investigations support the notion that source files associated with the same bug often share call relationships. This observation aligns with our initial premise stemming from the Class Name Match Score (Feature 1), suggesting that programming entities referenced in bug reports are likely to pinpoint the locus of bugs. It follows, then, that the scope of these entities should extend beyond individual files to encompass related files through call relationships.

In pursuit of this extended scope, we propagate the Class Name Match Score throughout the call graph. The Call Graph Score f_2 for a file s_k is thus calculated by aggregating the scores from its interconnected files:

$$f_2(s_k) = \sum_{s \in \text{Callers}(s_k)} f_1(s) + \sum_{s \in \text{Callees}(s_k)} f_1(s)$$

Here, $\text{Callers}(s)$ and $\text{Callees}(s)$ represent the sets of files that call and are called by s , respectively. The Class Name Match Score, $f_1(s)$, quantifies the relevance of each file s based on the mentioned class names.

This formula encapsulates the aggregated impact of both upstream and downstream dependencies, reflecting the compounding risk of bugs disseminated through the call graph. It thus offers a nuanced prediction of a file's susceptibility to bugs based on its contextual role within the software architecture.

Employing the Call Graph Score enables a multidimensional analysis that fuses direct textual correlations with the structural

dynamics of software interactions. The result is a holistic and nuanced assessment of each file's potential faultiness, enhancing the precision of the fault localization process.

3.4.3 Features from Existing Work. We incorporate five features from previous studies [38] that also utilize learning-to-rank for IRFL.

Text Similarity Score: The similarity between a bug report r and a source file s is quantified using the cosine similarity measure, defined as the cosine of the angle between their respective tf-idf vectors. This metric is particularly adept at highlighting textual relationships, ranging from explicit term matches to more subtle linguistic correlations.

API-Enriched Lexical Similarity: This feature measures the similarity between a bug report and the technical terms found in source code APIs. It is defined as:

$$f_4(r, s) = \max\{\text{sim}(r, s.\text{api}), \text{sim}(r, m.\text{api}) \mid m \in s\}$$

where sim is the similarity measure between the bug report r and the API documentation for the source file s or its methods m .

Collaborative Filtering Score: This score evaluates the similarity between a bug report and historical reports of the same source file, leveraging past bug fix data:

$$f_5(r, s) = \text{sim}(r, \text{br}(r, s))$$

Bug-Fixing Recency: This feature reflects the time since the last bug fix, calculated as:

$$f_6(r, s) = \frac{1}{(r.\text{month} - \text{last}(r, s).\text{month} + 1)}$$

More recent fixes yield a higher score.

Bug-Fixing Frequency: This metric counts the number of previous fixes for a file:

$$f_7(r, s) = |\text{br}(r, s)|$$

Files with more past fixes receive a higher score.

3.5 Feature Normalization

In preparation for training the ranking models, it is crucial to normalize the feature values to a consistent range, specifically $[0, 1]$. This normalization is conducted according to the following scheme:

$$f'_i = \begin{cases} 0 & \text{if } f_i < f_{i.\text{min}} \\ \frac{f_i - f_{i.\text{min}}}{f_{i.\text{max}} - f_{i.\text{min}}} & \text{if } f_{i.\text{min}} \leq f_i \leq f_{i.\text{max}} \\ 1 & \text{if } f_i > f_{i.\text{max}} \end{cases}$$

Here, f_i represents the original value, while f'_i denotes the normalized value of the i^{th} feature associated with a potentially buggy method. The terms min_i and max_i are the minimum and maximum values of the i^{th} feature, as determined from the training dataset.

3.6 Learning-to-Rank for Fault Localization

Learning-to-Rank (Ltr) models have revolutionized the field of information retrieval by providing a mechanism to learn complex ranking functions from data. In the domain of fault localization, an Ltr model can be trained to rank source files in order of their likelihood of containing bugs, as indicated by a given bug report.

SVM^{rank} is a specialized instance of Support Vector Machines (SVMs) tailored for ranking tasks, as proposed by T. Joachims. It operates by learning a linear ranking function that minimizes the average number of misordered pairs of documents (or in our case, source files). SVM^{rank} is chosen for its efficiency in handling sparse feature spaces and its effectiveness in learning with linear kernels, which is particularly beneficial given the high dimensionality typical of text data in software repositories. Additionally, SVM^{rank} 's capacity to process large datasets makes it well-suited for the vast amounts of source code and bug report data.

3.6.1 Training Process. These features are combined in a single feature vector to train the SVM^{rank} model, allowing for a comprehensive and nuanced ranking of source files according to their likelihood of containing bugs. To train our bug report classification model, we utilized a chronological cross-validation approach. Since different bug reports types have different characteristics, this involved dividing our dataset into 10 subsets for each categories (i.e., PE, ST, NL) based on chronological order, ranging from the oldest (Subset_1) to the most recent (Subset_{10}). We trained our model on earlier subsets and tested it on later subsets, mimicking real-world scenarios where models are trained on recent data and tested on future data. This process ensures that our bug report classification model is robust and generalizable, with evaluations conducted using standard metrics to assess performance and iteratively improve the model's accuracy and reliability.

3.6.2 Fault Localization. Our fault localization process involves giving a bug report to GPT-4, which will analyze it and generate queries based on our prompts. We then extract each feature for our trained learning-to-rank model. The model computes a score for each source file in the software project and uses this value to rank all the files in descending order. Users receive a ranked list of files, where higher-ranked files are more likely to be relevant to the bug report, indicating a higher likelihood of containing the bug's cause.

4 EVALUATION

We evaluate the effectiveness of queries generated by GPT-4 using various prompts to address the following research questions:

RQ1: How does the performance of LLmiRQ in IRFL compare to the state-of-the-art techniques? The goal of this research question is to assess the effectiveness of LLmiRQ in fault localization when compared to the state-of-the-art techniques.

RQ2: How does the effectiveness of our Ltr model compare to state-of-the-art IR models? The goal of this research question is to evaluate the effectiveness of each feature and LLmiRQ's performance across different retrieval models.

RQ3: How effective are our queries in IRFL? The goal of this research question is to evaluate the effectiveness of our query compared to existing query construction methods, the number of conversational-based reformulations, and the differences between 0-shot and 1-shot scenarios.

4.1 Experiment Setup

4.1.1 Dataset Collection. The dataset was collected by Lee et al. [18] and comprises 46 projects from Apache, Commons, JBoss, Spring,

Table 2: Dataset.

Projects	#PE	#ST	#NL	#BR	Projects	#PE	#ST	#NL	#BR
CAMEL	603	113	206	922	DATAREDIS	39	8	0	47
HBASE	334	117	59	510	DATAREST	77	16	15	108
HIVE	467	190	381	1038	LDAP	38	7	3	48
CODEC	38	2	1	41	MOBILE	11	0	0	11
COLLECTIONS	80	3	1	84	ROO	100	30	63	193
COMPRESS	84	17	8	109	SEC	259	42	20	321
CONFIGURATION	87	14	2	103	SECOAUTH	34	5	14	53
CRYPTO	1	0	0	1	SGF	52	17	10	79
CSV	12	1	1	14	SHDP	27	5	12	44
IO	71	8	2	81	SHL	2	0	6	8
LANG	147	15	1	163	SOCIAL	12	1	0	13
MATH	190	9	6	205	SOCIALFB	11	3	1	15
WEAVER	0	0	2	2	SOCIALLI	4	0	0	4
ENTESB	0	2	2	4	SOCIALTW	8	0	0	8
JBMETA	8	10	1	10	SPR	74	22	7	103
AMQP	59	18	13	90	SWF	69	2	11	82
ANDROID	4	2	2	8	SWS	79	25	6	110
BATCH	247	23	28	298	ELY	17	0	4	21
BATCHADM	11	2	5	18	SWARM	29	9	7	45
DATACMNS	100	28	8	136	WFARQ	1	0	0	1
DATAGRAPH	2	0	0	2	WFCORE	156	68	100	324
DATAJPA	99	27	10	136	WFLY	238	133	115	486
DATAMONGO	165	58	16	239	WFMP	2	0	0	2
	Total				4148	1043	1149	6340	

and Wildfly, with a total of 9,459 bug reports. The dataset is employed in a reproducibility study focused on the performance of IRFL. The size of this dataset is comparable to those utilized in existing IRFL studies [26, 28, 33, 35, 39, 41], which range from 3,479 to 5,139 bug reports. Typically, Users submit bug reports in relation to specific versions of a project. Regarding standard IRFL evaluation, a common approach involves a single version matching strategy, assuming the search encompasses potential source code files from a project’s most recent version. Our evaluation also adopts a single version matching strategy. However, due to the inclusion of some bug reports originating from very old project versions, corresponding source files are often unavailable. The absence of these source files introduces potential ambiguity and uncertainty into our evaluation process. As a solution, we opt to eliminate bug reports lacking corresponding fixed source files from the dataset. This process aims to ensure the quality and reliability of the dataset, leading to a more accurate representation of the data and enhancing the validity of our research results. After the process, the dataset comprises a total of 6,340 bug reports, as shown in Table 2.

4.1.2 Bug Reports Classification. As outlined in Section 3, the quality of bug reports is vital for IRFL effectiveness. Table 2 categorizes bug reports based on their content: **PE** denotes reports with programming entities in their descriptions. **ST** refers to reports containing stack traces. **NL** represents reports composed solely of plain natural language, excluding programming entities and stack traces. **BR** refers to the total number of bug reports. These categories are mutually exclusive; for instance, a report featuring stack traces is classified under ST regardless of the presence or absence of programming entities in its description.

4.1.3 IRFL Techniques. We select seven IRFL techniques to assess the effectiveness of LLmiRQ and LLmiRQ+ in IRFL tasks. Table 3 presents an overview of the settings for each technique. Each of these techniques considers various resources, which are generally classified into two categories: Bug Report and External Features. Bug Report resources encompass text descriptions (BRT) and stack traces (ST), while External Features resources encompass version

Table 3: Comparison of IRFL Techniques.

Technique	Bug Report		External Features			IR Model
	BRT	ST	VH	SB	SF	
BugLocator	•			•		rVSM
BLUiR	•			•	•	Indri
BRTtracer	•	•		•	•	rVSM
AmaLgam	•			•	•	Mixed
BLIA	•	•	•	•	•	rVSM
BLIZZARD	•	•			•	Lucene
LR	•		•	•		LtR

BRT: Bug Report Text, ST: Stack Trace, VH: Version History, SF: Source File, SB: Similar Bug Report

history (VH), similar bug reports (SB), and source files (SF). Specifically, **BugLocator** [41] suggests relevant files by examining similar bugs through rVSM. **BLUiR** [28] differentiates between summary and description in bug reports, employing distinct query and document representations for separate searches. **BRTtracer** [35] extends BugLocator by incorporating segmentation and stack-trace analysis to rank files. **AmaLgam** [33] employs three score components (version history, similar report, and structure) to rank files. **BLIA** [39] adapts the approaches of BugLocator, BLUiR, and BRTtracer, employing Google’s algorithm for version history analysis. **BLIZZARD** [26] transforms bug report text, stack traces, and source code into graphs, utilizing graph-based term weighting and Lucene for fault localization. **LR** [38] uses text similarity, API specification similarity, bug fix frequency, and bug fix recency as features, employing learning-to-rank (LtR) for training.

4.1.4 Performance Metrics. We use three performance metrics to evaluate our approach and other fault localization techniques. These metrics are widely used in the field and are also used by state-of-the-art techniques [26, 28, 33–35, 38, 39, 41].

Top@K measures the probability that the fault localization method will successfully locate the relevant bug reports when reviewing the first k files (where $k = 1, 5, 10$) in the recommendation list generated by the method. The calculation is $\text{Top}@K = \frac{|R_k|}{n}$, where $|R_k|$ is the total number of bug reports that are successfully located when the method is recommended by TopK. n is the total number of bug reports used in the evaluation process.

Mean Reciprocal Rank (**MRR**) [32] measures the position of the first source file related to the bug report located by the fault localization method in the recommendation list. The calculation is $\text{MRR} = \frac{1}{n} \sum_{j=1}^n \frac{1}{\text{rank}_j}$, where rank_j represents the ranking position of the first source files related to the j -th bug report in the recommendation list.

Mean Average Precision (**MAP**) [20] measures the average position of all source files related to the bug report located by the fault localization method in the recommendation list. The calculation is

$$\text{MAP} = \frac{1}{n} \sum_{j=1}^n \text{AvgP}_j, \quad \text{AvgP}_j = \frac{1}{|K_j|} \sum_{k \in K_j} \{\text{Prec}@k\}, \quad \text{Prec}@k = \frac{\sum_{i=1}^k \text{IsRelevant}(k)}{k},$$

where AvgP_j is the average precision for the j -th bug report, and $|K_j|$ is the total number of relevant source files for the j -th bug report. $\text{Prec}@k$ represents the precision of the top k files in the recommendation list, and $\text{IsRelevant}(i)$ returns 1 if the i -th file in the recommendation list is related to the bug report, and 0 otherwise. K_j is the true positive result set of a query. The higher value of each metric represents better performance.

Table 4: Comparing with State-of-the-Art Techniques.

	Tech	Top1	Top5	Top10	MRR	MAP
PE	BugLocator	50.95%	75.92%	82.48%	0.6149	0.4657
	BLUIR	42.19%	70.75%	78.68%	0.5402	0.4263
	BRTtracer	55.13%	77.87%	84.46%	0.6508	0.4875
	AmaLgam	42.51%	71.04%	79.02%	0.5432	0.4293
	BLIA	47.01%	74.90%	83.36%	0.5886	0.4632
	BLIZZARD	40.08%	66.54%	74.90%	0.5128	0.3833
	LR	48.38%	72.57%	81.01%	0.5860	0.4263
ST	LLmiRQ	61.91%	81.54%	86.12%	0.7042	0.5298
	LLmiRQ+	64.16%	83.79%	88.37%	0.7265	0.5524
	BugLocator	39.66%	68.34%	77.93%	0.5183	0.3954
	BLUIR	29.21%	55.86%	65.99%	0.4065	0.3073
	BRTtracer	46.16%	76.12%	84.22%	0.5897	0.4392
	AmaLgam	29.32%	55.97%	66.31%	0.4076	0.3084
	BLIA	46.06%	74.20%	83.26%	0.5797	0.4574
NL	BLIZZARD	36.46%	60.98%	69.72%	0.4663	0.3486
	LR	32.62%	62.37%	75.16%	0.4513	0.3440
	LLmiRQ	51.71%	79.64%	85.61%	0.6391	0.4802
	LLmiRQ+	52.45%	80.38%	86.35%	0.6461	0.4872
	BugLocator	24.66%	51.16%	64.80%	0.3632	0.2560
	BLUIR	24.95%	52.42%	62.96%	0.3668	0.2615
	BRTtracer	28.05%	54.84%	67.21%	0.3967	0.2736
ALL	AmaLgam	25.05%	52.51%	63.15%	0.3679	0.2618
	BLIA	17.50%	45.94%	58.12%	0.2875	0.2032
	BLIZZARD	21.47%	43.91%	55.22%	0.3133	0.2117
	LR	23.99%	60.13%	76.65%	0.3507	0.2575
	LLmiRQ	32.24%	61.32%	72.63%	0.4608	0.3228
	LLmiRQ+	37.43%	64.51%	75.82%	0.4923	0.3540
	BugLocator	44.33%	70.18%	78.53%	0.5534	0.4161
ALL	BLUIR	36.93%	64.98%	73.74%	0.4868	0.3769
	BRTtracer	48.75%	73.41%	81.30%	0.5947	0.4408
	AmaLgam	37.18%	65.21%	74.06%	0.4892	0.3791
	BLIA	41.51%	69.54%	78.77%	0.5326	0.4151
	BLIZZARD	36.11%	61.52%	70.48%	0.4690	0.3465
	LR	40.96%	67.62%	77.97%	0.5212	0.3865
	LLmiRQ	55.21%	77.56%	83.59%	0.6494	0.4841
ALL	LLmiRQ+	57.98%	80.33%	86.36%	0.6770	0.5118

4.2 Results and Analysis

4.2.1 RQ1: How does the performance of LLmiRQ in IRFL compare to the state-of-the-art techniques? To answer this RQ, we conducted an evaluation of LLmiRQ and LLmiRQ+ by comparing them with seven state-of-the-art IRFL techniques on our testing set, i.e., (*Subset₂*) to (*Subset₁₀*). To ensure fairness in the evaluation, we used the same configuration and parameters specified in each paper for all the techniques. Table 4 presents the evaluation results of LLmiRQ and LLmiRQ+. The technique with the best performance on each metric is boldfaced. The result marked with “**” indicates the technique is not statistically distinguishable from the best algorithm at $p = 0.05$ using paired t-tests. Results without being boldfaced or “**” mean they are significantly lower than the best performance. On average, LLmiRQ surpasses other methods across all five metrics, achieving a 0.6494 score on MRR and a 0.4841 score on MAP, besting the top-performing IRFL technique, BRTtracer, by 9% and 10% respectively. Especially, compared to the existing learning-to-rank method LR, LLmiRQ outperformed it by 25% on both MRR and MAP.

For PE (Programming Entities), considered ideal for IRFL, all techniques demonstrate better performance compared to ST and NL. In particular, BRTtracer achieves a high MRR of 0.6508 and a high MAP of 0.4875. However, LLmiRQ surpasses BRTtracer with a Top1 accuracy of 61.91%, Top5 accuracy of 81.54%, Top10 accuracy of 86.12%, an MRR of 0.7042, and a MAP of 0.5298, outperforming BRTtracer’s MRR and MAP by 8% and 9%, respectively. These results indicate that LLmiRQ effectively reduces noise and enhances the performance of IRFL.

For bug reports that include stack traces (ST), BRTracer introduced a scoring method based on the rank position of the entries in the stack traces. BLIA has since adopted this method from BR-Tracer. As the results indicate, this approach effectively computes the scores, leading their IRFL performance to surpass other state-of-the-art techniques. LLmiRQ leverages GPT-4’s strength which lies in its contextual comprehension to analyze stack traces. By analyzing the entire stack trace, the model can discern patterns or anomalies. For instance, it can identify if a specific method call consistently precedes an error, suggesting potential trouble spots. For ST, LLmiRQ achieves an MRR and MAP of 0.6391 and 0.4802, outpacing BRTtracer by 8% and 9%, respectively, and outpacing BLIA by 10% and 5%, respectively.

For bug reports described only in plain text without programming entities or stack traces (NL) (Table 4), BLIZZARD proposed an expansion method via pseudo-relevance feedback; other state-of-the-art techniques lack specific methods. Results suggest that while BLIZZARD’s expansion may be beneficial in certain scenarios, on average, it is less efficient than methods treating bug reports as queries alone. This inefficiency may arise from improper expansions introducing extraneous information. LLmiRQ leverages GPT-4’s expansive knowledge and contextual understanding to expand queries. LLmiRQ achieves performance metrics of 32.24% for Top1, 61.32% for Top5, 72.63% for Top10, an MRR of 0.4608, and a MAP of 0.3228 on *NL*, surpassing all competitors. This suggests that LLmiRQ adeptly infers and recommends classes or components associated with the given bug report.

After the performance evaluation of LLmiRQ, it was observed that 986 out of 6,340 bug reports did not rank in the generated list. This discrepancy can be attributed to the inadequate or misleading information contained within these reports, such as incorrect references to non-buggy classes. To address this issue, LLmiRQ+ was employed to perform query reformulation on these problematic bug reports, aiming to refine the queries and improve the accuracy of the retrieval results. After the reformulation process executed by LLmiRQ+, we observed that 156 out of the 986 previously unranked bug reports were successfully included in the ranking list. This improvement indicates that the query reformulation capability of LLmiRQ+ can effectively rectify issues related to inadequate or misleading information in bug reports, thereby enhancing the overall accuracy and reliability of the fault localization process.

RQ1 summary: The evaluation results demonstrate that LLmiRQ and its advanced version, LLmiRQ+, surpass six leading IRFL techniques. LLmiRQ+ consistently delivers exceptional results, outdoing BRTtracer by 9% in MRR and 10% in MAP. LLmiRQ also excels in analyzing stack traces (ST), leveraging GPT-4’s contextual understanding to outperform competing techniques. In handling plain text descriptions (NL), LLmiRQ shows remarkable proficiency, effectively inferring and recommending relevant components. With LLmiRQ+, which incorporates dynamic user-driven query reformulation, the system further enhances its performance, overcoming real-world challenges and surpassing existing state-of-the-art solutions.

4.2.2 RQ2: How does the effectiveness of our Ltr model compare to state-of-the-art IR models? To answer this research question, we first evaluate how each feature impacts our Learning-to-Rank model and determine the best combination of features. Then, we compare our optimized Learning-to-Rank model with state-of-the-art retrieval models.

Features Selection. With a set of 7 features (Section 3.4), we conducted a comprehensive evaluation of all possible combinations to ascertain the most effective feature set. Table 5 presents the outcomes, highlighting the optimal feature combinations tailored to each category of bug reports. Given that our queries may not always align with class names and method names, these features could sometimes result in zero scores. Similarly, call graph and class hierarchy features and other features may not always be applicable, but textual similarity between bug reports and source files generally provides a consistent score. Thus, textual similarity serves as a foundational feature, complemented by the other features.

Table 5: Comparing with Different Features.

Features	PE		ST		NL	
	MRR	MAP	MRR	MAP	MRR	MAP
TS	0.5785	0.4356	0.5316	0.3997	0.3273	0.2340
TS+CL	0.6654	0.4992	0.6055	0.4454	0.4124	0.3127
TS+CL+CG	0.6749	0.5056	0.6333	0.4644	0.3528	0.2604
ALL	0.6686	0.5014	0.6152	0.4537	0.3528	0.2604

TS: Text Similarity, CL: Class Match, CG: Call Graph, ALL: All features combined.

The results show that for bug reports with programming entities (PE) and stack traces (ST), the queries yield accurate results, with class match and call graph scores proving highly effective. The combination of text similarity (TS), class match (CL), and call graph (CG) delivers the best performance, while adding more features can increase false positives and reduce effectiveness. For natural language (NL) reports, LLmiRQ's GPT-4-driven queries suggest multiple root causes, mixing true and false positives. In this case, combining class matching with textual similarity works best, as incorporating call graph scores tends to lower overall accuracy due to increased false positives.

Comparison with State-of-the-Art Retrieval Models. We evaluated our learning-to-rank model against leading retrieval models, including BugLocator, BLUiR, BRTtracer, AmaLgam, BLIA, and BLIZZARD, using a randomized subset of the dataset and focusing on MRR and MAP. The results, shown in Figure 2a, highlight our model's superior performance and its alignment with the specialized nature of our queries, demonstrating its suitability for retrieval tasks.

RQ2 summary: We identified the best feature combinations for different bug report categories. For PE and ST, the optimal mix included Text Similarity, Class Match, and Call Graph, while for NL reports, Text Similarity and Class Match were sufficient. Our Learning-to-Rank model outperformed traditional IR models, demonstrating its effectiveness and superior performance.

4.2.3 RQ3: How effective are our queries in IRFL? To answer to this question, we compared our queries against others commonly used in IRFL, then examined the effects of conversational query reformulation and the influence of one-shot/zero-shot learning on our query performance.

Compared with Different Queries. To determine the effectiveness of our queries, we benchmarked them against common IRFL practices, which use entire bug reports (BR) as queries, and the BLIZZARD approach, which generates queries using a graph-based method. To ensure consistency in our evaluation, we perform all queries using our learning-to-rank model. As shown in Table 6, our queries surpass both BR and BLIZZARD for PE and ST categories, signifying that our method enables GPT-4 to more accurately pinpoint the true root causes within bug reports. For NL categories, despite BLIZZARD's use of Pseudo-relevance feedback, our queries still perform better, demonstrating the efficacy of leveraging GPT-4's knowledge to refine the queries.

Table 6: Comparing with Different Queries.

Features	PE		ST		NL	
	MRR	MAP	MRR	MAP	MRR	MAP
BR	0.5172	0.3804	0.3931	0.2702	0.3599	0.2051
BLIZZARD	0.5161	0.3814	0.4018	0.2765	0.3660	0.2574
LLmiRQ	0.7680	0.5816	0.6554	0.4841	0.4867	0.3030

Conversation Cycles. To assess the efficiency of our query reformulation process, we analyzed 1,743 bug reports from Dataset requiring reformulation. Setting a limit of 5 conversation cycles, we recorded MRR and MAP performance. Figure 2b reveals that peak performance is typically reached in the first cycle, averaging an MRR of 0.3584 and a MAP of 0.2428. Our conversation-based reformulation efficiently responds to initial user feedback, resulting in significant query improvement. Subsequent cycles provide marginal enhancements, suggesting rapid peak performance attainment. Hence, we implemented LLmiRQ+'s one conversation cycle.

0/1 Shot. 0/1-shot learning assesses the model's adaptability across diverse tasks without extensive fine-tuning. Our experiment, conducted on a randomized dataset subset, yielded results in Table 7. The 1-shot approach outperformed 0-shot in PE, ST, and NL bug report categories. While 1-shot showed no significant difference from 0-shot in extracting programming entities, tasks in ST and NL resembled 0-shot for GPT. Notably, the 1-shot method excelled in ST and NL, leveraging initial context for accurate queries. The 0-shot occasionally produced non-optimally formatted output, requiring human effort to reformat queries. Consequently, we adopted the 1-shot approach for LLmiRQ and LLmiRQ+.

RQ3 summary: Our queries outperformed standard IRFL queries in PE and ST categories and remained effective in NL categories. Conversational query reformulation led to quick performance gains, with one-shot learning proving especially beneficial for ST and NL bug reports.

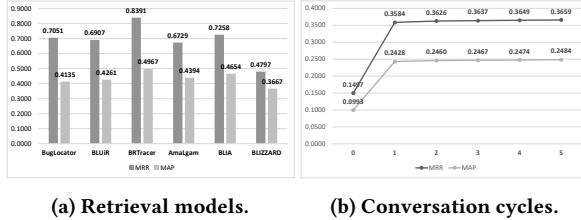


Figure 2: The impact of various components on LLmiRQ.

Table 7: Comparison between 0-shot and 1-shot.

	shots	Top1	Top5	Top10	MRR	MAP
PE	0-shot	60.32%	82.02%*	84.79%*	0.6958	0.4999
	1-shot	61.17%*	81.38%	83.51%	0.7014*	0.5024*
ST	0-shot	34.34%	78.85%	87.91%	0.5241	0.4044
	1-shot	59.62%	84.89%	87.91%	0.7092	0.5545
NL	0-shot	14.89%	37.81%	48.48%	0.2453	0.1567
	1-shot	20.16%	47.96%	57.71%	0.3158	0.2020
ALL	0-shot	28.43%	47.49%	52.44%	0.3641	0.2599
	1-shot	35.17%	56.07%	61.33%	0.4423	0.3148

5 RELATED WORK

Large Language Model. In recent years, Large Language Models (LLMs) [11, 24, 31] have significantly advanced the field of natural language processing by offering transformative capabilities in understanding and generating human-like text. LLMs, such as OpenAI’s GPT [22] (Generative Pre-trained Transformer) series, are trained on extensive corpora of textual data and utilize deep learning techniques, particularly transformer architectures, to capture intricate patterns and nuances in language. Their applications span a wide range of tasks, from text completion and translation to more complex endeavors like question-answering, summarization, and even code generation.

Recently, LLMs have been widely applied to software engineering tasks, such as code generation, code completion [8, 12, 15], code search [29, 30], and program repair [36, 40]. Ciborowska et al. [9] presented an approach for automatically retrieving bug-inducing changesets for newly reported bugs, using the popular BERT model to more accurately match the semantics in the bug report text to the inducing changeset. They also proposed using data augmentation (DA) [10] to create new, realistically looking bug reports that can significantly increase the size of the training set.

More recently, Kang et al. [17] proposed AutoFL, a fault localization method based on LLMs that generates both fault locations and natural language explanations for bug reports. By effectively leveraging LLMs to provide rationales alongside fault predictions, AutoFL improves developer trust in the recommendations. Additionally, Yoo et al. [23] introduced AgentFL, a multi-agent fault localization system built on ChatGPT. By simulating a three-step debugging process (comprehension, navigation, and confirmation), AgentFL employs diverse agents with specialized expertise and utilizes strategies such as test behavior tracking and multi-round dialogue to localize faults effectively.

However, these works do not focus on IR-based fault localization. Ciborowska’s work requires training a model, while AutoFL

and AgentFL rely on test coverage information. In contrast, our approach only relies on bug reports and simply prompts the model.

Information Retrieval-based Fault Localization. IR-based fault localization is an effective technique for identifying bugs in software systems. Automated methods are particularly valuable in this context to circumvent the time-consuming and error-prone nature of manual code inspection. Gay et al. [14] proposed a relevance feedback mechanism to enhance the vector space model (VSM) used in retrieval processes. Zhou et al. [41] further improved the VSM model by incorporating document length, leading to the development of the revised VSM (rVSM). Saha et al. [28] utilized the Eclipse JDT to parse the source code’s abstract syntax tree (AST) and extract information from four types of code entities, thereby augmenting the efficiency of IR-based fault localization. Additionally, Wang et al. [33] introduced an approach that leverages version history information to refine IR-based fault localization techniques. In our research, we have conducted a comparative analysis of our approach with these established localization techniques. The results demonstrate that our approach can outperform all these methods, showcasing its effectiveness in the realm of fault localization.

6 THREATS TO VALIDITY

The primary threat to the external validity of this study involves its generalizability across domains, programming languages, project sizes, and user expertise levels. To address these concerns, we employed a dataset comprising 46 diverse projects, enhancing the approach’s applicability. We also mitigated user expertise variability by implementing a structured feedback template, minimizing the reliance on users’ technical articulation and familiarity with the software system.

The primary threat to internal validity stems from reliance on the GPT model for bug report analysis, which may be limited by its understanding of complex technical contexts. The accuracy of generated queries and the effectiveness of query reformulation hinge on GPT’s interpretation of user feedback, introducing the risk of inaccurate adjustments due to misinterpretation or ambiguity. To address these concerns, we emphasize GPT’s advanced capabilities and proven efficacy in diverse applications. Recognized as one of the most powerful language models, GPT has demonstrated robust performance across numerous domains, accurately interpreting and processing complex language constructs in various tasks.

7 CONCLUSION

Our approach, named LLmiRQ, categorizes bug reports into programming entities, stack traces, and pure text, tailoring query construction strategies. Introducing LLmiRQ+, a novel user feedback-driven approach. Moreover, to optimize query effectiveness further, we have incorporated a learning-to-rank model that utilizes crucial features including class name match scores and call graph scores. We demonstrated significant outperformance of existing IRFL techniques in our evaluation of 46 projects and 6,340 bug reports, achieving an MRR of 0.6770 and a MAP of 0.5118. This marks a notable advancement in software engineering. Future work will refine GPT-based fault localization, with a focus on enhancing method-level capabilities.

REFERENCES

- [1] [n. d.]. Bugzilla. <https://bugzilla.mozilla.org/describekeywords.cgi>.
- [2] [n. d.]. GitHub. <https://github.com>.
- [3] [n. d.]. gooledoc. <https://code.google.com>.
- [4] Tien-Duy B. Le, David Lo, Claire Le Goues, and Lars Grunske. 2016. A learning-to-rank based fault localization approach using likely invariants. In *Proceedings of the 25th international symposium on software testing and analysis*. 177–188.
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [6] Oscar Chaparro, Juan Manuel Florez, and Andrian Marcus. 2017. Using observed behavior to reformulate queries during text retrieval-based bug localization. In *2017 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 376–387.
- [7] Oscar Chaparro, Juan Manuel Florez, and Andrian Marcus. 2019. Using bug descriptions to reformulate queries during text-retrieval-based bug localization. *Empirical Software Engineering* 24 (2019), 2947–3007.
- [8] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
- [9] Agnieszka Ciborowska and Kostadin Damevski. 2022. Fast changeset-based bug localization with BERT. In *Proceedings of the 44th International Conference on Software Engineering*. 946–957.
- [10] Agnieszka Ciborowska and Kostadin Damevski. 2023. Too Few Bug Reports? Exploring Data Augmentation for Improved Changeset-based Bug Localization. *arXiv preprint arXiv:2305.16430* (2023).
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [12] Jean-Baptiste Döderlein, Mathieu Acher, Djamel Eddine Khelladi, and Benoit Combemale. 2022. Piloting Copilot and Codex: Hot Temperature, Cold Prompts, or Black Magic? *arXiv preprint arXiv:2210.14699* (2022).
- [13] Juan Manuel Florez, Oscar Chaparro, Christoph Treude, and Andrian Marcus. 2021. Combining query reduction and expansion for text-retrieval-based bug localization. In *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 166–176.
- [14] Gregory Gay, Sonia Haiduc, Andrian Marcus, and Tim Menzies. 2009. On the use of relevance feedback in IR-based concept location. In *2009 IEEE international conference on software maintenance*. IEEE, 351–360.
- [15] Malihah Izadi, Roberta Gismondi, and Georgios Gousios. 2022. Codefill: Multi-token code completion by jointly learning from structure and naming sequences. In *Proceedings of the 44th International Conference on Software Engineering*. 401–412.
- [16] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 133–142.
- [17] Sungmin Kang, Gabin An, and Shin Yoo. 2024. A quantitative and qualitative evaluation of LLM-based explainable fault localization. *Proceedings of the ACM on Software Engineering* 1, FSE (2024), 1424–1446.
- [18] Jaekwon Lee, Dongsun Kim, Tegawendé F Bissyandé, Woosung Jung, and Yves Le Traon. 2018. Bench4bl: reproducibility study on the performance of ir-based bug localization. In *Proceedings of the 27th ACM SIGSOFT international symposium on software testing and analysis*. 61–72.
- [19] Zhengliang Li, Zhiwei Jiang, Xiang Chen, Kaibo Cao, and Qing Gu. 2021. Labrop: a label propagation-based software bug localization method. *Information and Software Technology* 130 (2021), 106410.
- [20] Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing.
- [21] Laura Moreno, John Joseph Treadway, Andrian Marcus, and Wuwei Shen. 2014. On the use of stack traces to improve text retrieval-based bug localization. In *2014 IEEE International Conference on Software Maintenance and Evolution*. IEEE, 151–160.
- [22] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023). <https://doi.org/10.48550/ARXIV.2303.08774> arXiv:2303.08774
- [23] Yihao Qin, Shangwen Wang, Yiling Lou, Jinhao Dong, Kaixin Wang, Xiaoling Li, and Xiaoguang Mao. 2024. AgentFL: Scaling LLM-based Fault Localization to Project-Level Context. *arXiv preprint arXiv:2403.16362* (2024).
- [24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [25] Mohammad Masudur Rahman and Chanchal K Roy. 2018. Effective reformulation of query for code search using crowdsourced knowledge and extra-large data analytics. In *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 473–484.
- [26] Mohammad Masudur Rahman and Chanchal K Roy. 2018. Improving ir-based bug localization with context-aware query reformulation. 621–632.
- [27] Peter C Rigby and Martin P Robillard. 2013. Discovering essential code elements in informal documentation. In *2013 35th International Conference on Software Engineering (ICSE)*. IEEE, 832–841.
- [28] Ripon K Saha, Matthew Lease, Sarfraz Khurshid, and Dewayne E Perry. 2013. Improving bug localization using structured information retrieval. In *2013 28th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 345–355.
- [29] Pasquale Salza, Christoph Schwizer, Jian Gu, and Harald C Gall. 2022. On the effectiveness of transfer learning for code search. *IEEE Transactions on Software Engineering* (2022).
- [30] Zejian Shi, Yun Xiong, Xiaolong Zhang, Yao Zhang, Shanshan Li, and Yangyong Zhu. 2022. Cross-Modal Contrastive Learning for Code Search. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 94–105.
- [31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [32] Ellen M. Voorhees. 1999. The TREC-8 Question Answering Track Report. In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999 (NIST Special Publication, Vol. 500-246)*, Ellen M. Voorhees and Donna K. Harman (Eds.). National Institute of Standards and Technology (NIST).
- [33] Shaowei Wang and David Lo. 2014. Version history, similar report, and structure: Putting them together for improved bug localization. In *Proceedings of the 22nd International Conference on Program Comprehension*. 53–63.
- [34] Ming Wen, Rongxin Wu, and Shing-Chi Cheung. 2016. Locus: Locating bugs from software changes. In *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering*. 262–273.
- [35] Chu-Pan Wong, Yingfei Xiong, Hongyu Zhang, Dan Hao, Lu Zhang, and Hong Mei. 2014. Boosting bug-report-oriented fault localization with segmentation and stack-trace analysis. In *2014 IEEE international conference on software maintenance and evolution*. IEEE, 181–190.
- [36] Chunqiu Steven Xia and Lingming Zhang. 2023. Conversational automated program repair. *arXiv preprint arXiv:2301.13246* (2023).
- [37] Chunqiu Steven Xia and Lingming Zhang. 2023. Keep the Conversation Going: Fixing 162 out of 337 bugs for \$0.42 each using ChatGPT. *arXiv preprint arXiv:2304.00385* (2023).
- [38] Xin Ye, Razvan Bunescu, and Chang Liu. 2014. Learning to rank relevant files for bug reports using domain knowledge. In *Proceedings of the 22nd ACM SIGSOFT international symposium on foundations of software engineering*. 689–699.
- [39] Klaus Changsun Youm, June Ahn, and Eunseok Lee. 2017. Improved bug localization based on code change histories and bug reports. *Information and Software Technology* 82 (2017), 177–192.
- [40] Wei Yuan, Quanjun Zhang, Tieke He, Chunrong Fang, Nguyen Quoc Viet Hung, Xiaodong Hao, and Hongzhi Yin. 2022. CIRCLE: Continual repair across programming languages. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*. 678–690.
- [41] Jian Zhou, Hongyu Zhang, and David Lo. 2012. Where should the bugs be fixed? more accurate information retrieval-based bug localization based on bug reports. In *2012 34th International conference on software engineering (ICSE)*. IEEE, 14–24.