

Noname manuscript No.  
(will be inserted by the editor)

# CodeFlowLM: Incremental Just-In-Time Defect Prediction with Pretrained Language Models and Exploratory Insights into Defect Localization

Monique Louise Monteiro · George G. Cabral · Adriano L. I. OLiveira

Received: date / Accepted: date

**Abstract** This work introduces CodeFlowLM, an incremental learning framework for Just-In-Time Software Defect Prediction (JIT-SDP) that leverages pre-trained language models (PLMs). Unlike traditional online learners, CodeFlowLM employs continual fine-tuning to address concept drift, class imbalance, and verification latency without retraining from scratch. We evaluated encoder-only and encoder-decoder PLMs (notably CodeT5+ and UniX-Coder) in JIT-SDP scenarios within and between projects, comparing them with the incremental baseline BORB. The results show that CodeFlowLM achieves up to 68% G-Mean gains, confirming its superior adaptability and robustness in evolving software environments. We further extend the analysis to Just-in-Time Defect Localization (JIT-DL), benchmarking Large Language Models (LLMs) such as GPT-5, Claude Sonnet 4.5, and Gemini 2.5 Pro against attention-based models. GPT-5 delivers comparable performance for Recall@20% and Effort@20% with higher stability, although attention-based methods retain an advantage in fine-grained ranking metrics (Top-k, IFA). A qualitative error analysis reveals that most false positives arise from (1) human-like conservative bias, (2) insufficient contextual information in diff-based prompts, and (3) potential dataset mislabeling in JIT-Defects4J. These findings highlight both the promise and the current limitations of LLM reasoning in defect localization. False negatives occur in smaller proportions. Overall, CodeFlowLM significantly advances the state of the art in incremental JIT-SDP, demonstrating superior adaptability and robustness in evolving software

---

Monique Louise Monteiro  
Av. Jornalista Aníbal Fernandes, s/n – Cidade Universitária. Recife-PE – Brazil  
Tel.: +55-81-2126-8430  
E-mail: mlbm@cin.ufpe.br

George G. Cabral  
Rua Dom Manuel de Medeiros, s/n, Dois Irmãos, Recife/PE – Brazil

Adriano L. I. OLiveira  
Av. Jornalista Aníbal Fernandes, s/n – Cidade Universitária, Recife-PE – Brazil

environments. Furthermore, our exploratory analysis of LLMs in JIT-DL not only benchmarks their performance against established attention-based models but also provides critical insights into the current limitations of prompt-based defect reasoning.

**Keywords** Just-in-Time Defect Prediction · Defect Localization · Pretrained Language Models · Incremental Learning · Large Language Models

## 1 Introduction

Just-in-Time Software Defect Prediction (JIT-SDP) is a software engineering paradigm designed to identify defects at the point of introduction, thus facilitating immediate remediation and lowering the costs associated with quality assurance. Contemporary research frames JIT-SDP as a multimodal problem (Ni et al., 2022a; Liu et al., 2024; Chen et al., 2024; Abu Talib et al., 2024), integrating various sources of information such as source code modifications, commit logs, and hand-crafted expert features (Liu et al., 2024; Chen et al., 2024; Kamei et al., 2013). Although early efforts relied on traditional machine learning and deep neural architectures, recent breakthroughs have been driven by the adoption of pre-trained language models, particularly those trained on large-scale code corpora, which have shown substantial gains in predictive accuracy (Liu et al., 2024; Abu Talib et al., 2024; Guo et al., 2023; Wang et al., 2024). These models vary not only in their design, encompassing encoder-based, decoder-based, and encoder-decoder frameworks, but also in their application, supporting both zero- and few-shot prompting, as well as domain-specific fine-tuning. Within the transformer (Vaswani et al., 2017) family, encoder modules specialize in capturing contextual representations of the input, whereas decoder modules are responsible for generating outputs conditioned on this context.

Encoder-oriented architectures such as CodeBERT (Feng et al., 2020) and UniXCoder (Guo et al., 2022) have shown strong capabilities in program comprehension, whereas decoder-only large language models (LLMs) tend to excel in tasks requiring complex reasoning. Despite these advances, comparative investigations of these model families within the context of JIT-SDP, particularly with respect to online or incremental learning, are virtually non-existent, even though real-world JIT-SDP environments naturally involve concept drift, evolving class imbalance, and verification latency (Cabral et al., 2019, 2023). Crucially, for Large Language Models (LLMs), a formal comparison against established encoder-only architectures remains largely unexplored even in the simpler, static (offline) JIT-SDP setting.

This work investigates the effectiveness of pre-trained Small Language Models (SLMs) in the context of within-project (WP) and cross-project (CP) JIT-SDP with incremental learning, as well as the generative capabilities of Large Language Models (LLMs) for the task of Just-in-Time Defect Localization (JIT-DL).

In regard to JIT-DL, while LLMs have recently demonstrated promising capabilities in broader fault localization and program repair tasks, evaluations have remained predominantly quantitative. This creates a critical blind spot: aggregate performance metrics alone cannot reveal why LLMs fail, whether their mistakes follow systematic patterns, or whether they are influenced by dataset noise, conservative heuristics, or missing context. A lack of qualitative insight limits both scientific understanding and practical deployment, as software engineering practitioners require predictable, interpretable behavior from automated tools.

Our investigation is guided by the following research questions:

- RQ1: **How do pre-trained language models perform in comparison to traditional machine learning approaches for continual within-project and cross-project Just-in-Time Software Defect Prediction (JIT-SDP)?** This RQ examines Small Language Models (SLMs) such as CodeT5+ and UniXCoder in incremental learning settings, comparing them with a state-of-the-art baseline (Cabral et al., 2023) for continual Just-in-Time Defect Prediction, introducing CodeFlowLM as a new framework for incremental JIT-SDP with pre-trained code language models. We investigated both within-project and cross-project performance.
- RQ2: **How do Large Language Models compare with attention-based mechanisms for Just-in-Time Defect Localization (JIT-DL)?** This research question addresses the problem of fault localization (FL) specifically in the context of just-in-time defect detection. Recent solutions are based on common techniques such as 1) attention-based fault localization (Ni et al., 2022a; Huang et al., 2024; Ju et al., 2025) or 2) multitask learning in both classifying commits in clean/defect-inducing and pointing out the possible bug locations in tokens or lines (Chen et al., 2024). We intend to investigate the capabilities of Large Language Models (LLMs) in this task from a quantitative perspective, as these models have recently been successfully applied in broader contexts, including file- and method-based software defect prediction and Automatic Program Repair (APR).
- RQ3: **Despite their near-human capabilities, in which situations do Large Language Models still underperform in fault localization?** This RQ presents the results of a qualitative analysis aimed at identifying common situations in which LLMs fail in defect localization, by randomly and manually inspecting samples of false positives and false negatives. Specifically regarding false positives, we hypothesize that these models are 1) biased towards conservative human behavior in bug localization, 2) confounded by a lack of sufficient context, and 3) that dataset noise may lead to answers mismatched with the supposedly ground truth. We believe that investigating these hypotheses can lead to improvements through prompt engineering as future work.

The main contributions of this work are as follows:

- C1 – CodeFlowLM: A novel incremental JIT-SDP framework based on PLMs, achieving 10–68% G-Mean improvements over BORB across most projects.
- C2 – First systematic evaluation of LLMs for JIT-DL, showing that GPT-5 achieves competitive Recall@20% and Effort@20% scores with higher stability.
- C3 – Comparative analysis of PLM attention-based defect localization, clarifying strengths and weaknesses relative to LLM reasoning.
- C4 – Qualitative error taxonomy for LLM-based localization, revealing conservative bias, context limitations, and labeling inconsistencies in the chosen evaluation dataset – JITDefects4J (Ni et al., 2022a).
- C5 – Empirical evidence that PLMs outperform traditional learners under continual learning, reinforcing their potential for real-world deployment.

This paper is structured as follows. Section 2 reviews related work, Section 3 details the experimental setup, Section 4 discusses results for each research question, and Section 5 concludes with key findings and future directions. The source code and the prompts will be publicly available on [https://github.com/monilouise/codewflowlm\\_jitdl](https://github.com/monilouise/codewflowlm_jitdl).

## 2 Related Work

### 2.1 Just-in-Time Software Defect Prediction

Shehab et al. (2024) address severe class imbalance in JIT-SDP by casting it as one-class classification trained only on normal commits, eliminating the need for oversampling/undersampling. Across 34 projects with cross-validation and time-aware evaluation, one-class SVM, Isolation Forest, and one-class k-NN consistently outperform binary SVM / RF / k-NN for medium to high imbalance, while using fewer features, reducing computation, and improving interpretability. Binary classifiers perform better only when the imbalance is low. The work positions OCC as a more efficient and scalable alternative for large, highly imbalanced software projects.

Li et al. (2024) systematically compare 10 sampling methods in 10 OSS projects (time-wise CV) to tackle class imbalance in JIT-SDP, showing that effectiveness depends on task and metric. For defect classification, Random Under-Sampling (RUM) yields the best F1, AUC, and MCC; sampling generally increases recall / F1 / AUC / MCC, but increases false alarms and lowers precision. They also find that window length (e.g., 2 vs. 6 months) has no consistent effect, and LR hyperparameter tuning offers negligible gains, so defaults are often sufficient, highlighting that sampling choice should align with specific evaluation goals.

None of the previous work uses semantic features (e.g., source code or commit messages) to represent the code changes. Instead, they rely only on 14 tabular/expert features proposed by Kamei et al. (2013).

## 2.2 Online and Incremental Just-in-Time Software Defect Prediction

Cabral et al. (2019) provide the first evidence that JIT-SDP operates under nonstationary class imbalance and severe verification latency, showing that defect rates fluctuate over time and that ignoring label delay leads to overly optimistic performance. They propose Oversampling Rate Boosting (ORB), an online latency-aware resampling strategy that dynamically adjusts the oversampling rate when predictions become skewed, while avoiding the amplification of noisy minority samples. Evaluated on ten GitHub projects, ORB achieves competitive G-mean and reduces the recall gap, establishing that handling imbalance evolution and latency is essential for robust JIT-SDP. Cabral et al. (2023) extend this line with BORB, a batch version of ORB that periodically retrains offline models on accumulated history while applying adaptive resampling at training and test times. On ten GitHub projects, BORB yields modest gains over ORB for several learners, and incorporating cross-project data boosts both methods – especially ORB – mitigating performance degradation. Although ORB is consistently more computationally efficient, BORB’s cost per day remains practical, clarifying the real-world trade-offs between accuracy and efficiency. Cabral and Minku (2023) investigate concept drift in JIT-SDP, analyzing shifts in  $p(y)$ ,  $p(x|y)$  and  $p(x)$  and showing that long verification latencies (days to more than 11 years) undermine the reliability of the classifier. They introduce PBSA, an online drift-adaptive method that monitors predictions on unlabeled data and adapts without needing immediate labels. In ten GitHub projects, PBSA achieves a higher and more stable performance than ORB, significantly reducing the recall gap and emerging as the most reliable JIT-SDP technique.

Song et al. (2023) propose HumLa, a realistic online labeling regime where developers immediately inspect commits predicted as defect-inducing rather than relying on delayed, retrospective labels. Rapid feedback enables faster adaptation, yielding markedly higher performance even under partial and noisy inspection. They also present ECo-HumLa, which prioritizes high-confidence predictions to reduce inspection costs while retaining accuracy comparable to full inspection, detecting more real defects and producing fewer false alarms than random inspection at equal effort. The results demonstrate that effort-aware and realistic labeling is vital for a reliable evaluation and deployment of JIT-SDP.

In the same way as in the previous subsection, none of the above works uses semantic features to represent the code changes, relying only on tabular/expert features. Furthermore, no work examines whether PLMs can be incrementally fine-tuned in a latency-aware streaming scenario, nor how they behave under drift and reclassification of commits. In contrast, our proposed CodeFlowLM is based on a combination of expert and semantic information.

### 2.3 Cross-Project Just-In-Time Software Defect Prediction

Tabassum et al. (2023) present the first online cross-project (CP) JIT-SDP study, introducing three continually updated strategies – All-in-One, ensemble, and filtering – that incrementally integrate both CP and within-project (WP) data streams. Evaluated on 9 proprietary and 10 open-source systems, their results show that combining CP+WP information consistently improves G-mean across all phases, and that online CP approaches outperform traditional offline baselines. They further propose an analysis framework to distinguish stable vs. drop performance periods and offer practical guidance regarding computational overhead and hyperparameter sensitivity in online JIT-SDP.

Zhu et al. (2024) address cross-project JIT-SDP with ISKMM, a method that reweights and selects source-project instances via Kernel Mean Matching to align them with the target distribution, followed by resampling to mitigate class imbalance. Tested on 10 projects using multiple learners, ISKMM surpasses CP single-source baselines, and CP models trained with ISKMM reach performance levels comparable to within-project predictors – demonstrating its effectiveness when local training data are limited or unavailable.

While cross-project strategies have progressed through weighting, filtering, and online self-adaptation, these methods again operate exclusively on expert features, ignoring code semantics and the capabilities of PLMs to transfer knowledge across repositories.

### 2.4 JIT-SDP and Deep Learning

DeepJIT (Hoang et al., 2019) is an end-to-end convolutional network for JIT defect prediction that learns directly from commit messages and code diffs, thus eliminating the need for handcrafted metrics and addressing class imbalance through unequal misclassification loss. In Qt and OpenStack, it achieves substantial AUC gains, with code changes contributing most to predictive power; moreover, adding manual features can further improve accuracy.

CC2Vec (Hoang et al., 2020) extends this line of work by learning hierarchical change representations (words → lines → hunks) with an attention mechanism comparing removed and added code, guided by commit-message semantics. The resulting language-agnostic embeddings benefit multiple software-engineering tasks and significantly increase the AUC of DeepJIT.

Zhou et al. (2024) show the value of combining deep semantic signals with expert knowledge. Their SimCom++ model integrates a Random Forest over handcrafted features (“Sim”) with a textCNN over commit logs and code changes (“Com”). Across six datasets, it consistently outperforms state-of-the-art baselines on ROC-AUC, PR-AUC, and F1, and analysis reveals that the two modules make complementary predictions that reduce false positives and correct errors — demonstrating the robustness gained from combining expert-crafted and deep-learned information for JIT-SDP.

It is important to highlight that Deep models such as CNNs and hierarchical diff encoders introduced semantic learning into JIT-SDP, but they were trained offline and cannot adapt to evolving data distributions, label delays, or concept drift. Moreover, these architectures predate modern PLMs and lack their capacity for large-scale semantic understanding.

#### *2.4.1 JIT-SDP and Pretrained Language Models*

Ni et al. (2022b) propose CompDefect, a multitask, function-level framework that unifies defect identification, defect type categorization, and automatic repair for JIT-SDP. It leverages GraphCodeBERT (Guo et al., 2021) to capture semantic and data-flow structure from clean, defect inducing, and fixed versions of a function, enabling more precise localization within commits. In benchmarks, CompDefect surpasses DeepJIT/CC2Vec for identification, BERT (Devlin et al., 2019)/RoBERTa (Liu et al., 2019)/CodeBERT (Feng et al., 2020) for categorization (F1), and Sequencer for repair (BLEU). The work addresses prior gaps – underuse of semantic/structural code signals and separated tasks – showcasing the benefits of joint modeling.

Ni et al. (2022a) propose JIT-Fine, a unified model for just-in-time defect prediction (JIT-DP) and defect localization (JIT-DL). It combines 14 change-level expert metrics with CodeBERT semantic features to learn joint representations that both flag defect inducing commits and pinpoint defect-inducing lines. To address dataset quality issues, they introduce JIT-Defects4J, a large, manually labeled, line-level dataset. Compared to six state-of-the-art baselines in ten metrics, JIT-Fine delivers substantial and consistent improvements on both tasks.

CCT5 (Lin et al., 2023) is a T5-based, code-change-oriented pre-trained model tailored to maintenance tasks that require an understanding of diffs. The authors build CodeChangeNet (more than 1.5M code-change/commit-message pairs across six languages) and pre-train with five objectives: masked modeling for code changes and commit messages, NL  $\leftrightarrow$  PL generation, and a structure-aware Code Diff Generation (CDG) using data flow. In downstream tasks – commit message generation, JIT comments, and JIT defect prediction – CCT5 achieves better results than the selected baselines.

Zhou et al. (2023) propose CCBERT, a Transformer pre-trained at token level on more than 1.3M Java code-change hunks that encode old/new code plus explicit edit actions (insert/delete/replace/equal). They introduce four self-supervised objectives to learn generic code-change representations without relying on noisy commit messages. Across downstream tasks (JIT defect prediction, patch correctness, bug-fixing commit prediction), CCBERT outperformed CC2Vec and even larger models (CodeBERT, GraphCodeBERT) while requiring less training/inference time and GPU memory.

Liu et al. (2024) deliver the first empirical study of PEFT for dynamic code-change tasks, comparing Adapter Tuning (Houlsby et al., 2019) and LoRA (Hu et al., 2022) to full-model fine-tuning and other baselines across five PLMs. For JIT-SDP, Adapter Tuning, and LoRA achieve state-of-the-art results –

especially when augmented with expert features – and prove robust in cross-lingual and low-resource settings. The work offers practical guidance on using PEFT to efficiently adapt models for code-change applications.

Guo et al. (2024) conduct the first systematic study of fine-tuning choices for BERT-style models (CodeBERT, RoBERTa) in JIT-DP. They find that the first encoder layer is pivotal – freezing it hurts performance – and adding AdamW yields small gains. Based on these insights, a cost-effective LoRA fine-tuning approach matches full fine-tuning with roughly one-third the memory, improving practicality for deployment.

#### *2.4.2 JIT-SDP and Large Language Models*

Kim et al. (2025) review JIT-DP from its early 2000s origins to today, diagnosing barriers to industrial uptake: unreliable SZZ labeling, verification latency, and deployment/maintenance complexity. They emphasize the importance of strong explainability to foster developer trust. Looking ahead, they argue that LLMs are a paradigm shift: offering application-level code understanding, fine-grained explanations, and fix suggestions — potentially unifying defect prediction, categorization, and repair.

Monteiro et al. (2025) present a broad empirical comparison of pre-trained code LMs for JIT-SDP, testing encoder-only, decoder-only, and encoder-decoder models (e.g., CodeT5+, UniXCoder, CodeReviewer) against closed API LLMs (GPT, Gemini) via fine-tuning and zero-shot prompting on JIT-Defects4J. Fine-tuned small to medium open models — especially CodeT5+ and UniXCoder — substantially outperformed zero-shot closed LLMs; encoder and encoder-decoder architectures suit this discriminative task better than decoder-only models. Combining semantic signals (diffs + messages) with expert features yielded the best results. CodeT5+ and UniXCoder achieved SOTA in cross-project JIT-SDP, surpassing previous baselines like JIT-Smart. This was the first direct comparison of open (trainable) vs. closed (prompt-based) decoder-only models for JIT-SDP.

### 2.5 Defect Localization

#### *2.5.1 Just-in-Time Software Defect Localization*

Pornprasit and Tantithamthavorn (2021) proposed JITLine, a lightweight JIT defect predictor designed to overcome key shortcomings of deep models such as CC2Vec and DeepJIT. By combining Bag-of-Tokens features, optimized SMOTE (Chawla et al., 2002), and LIME (Ribeiro et al., 2016) for line-level interpretation, JITLine achieves faster predictions and higher commit-level accuracy in OpenStack and Qt, outperforming deep and NLP baselines while reducing triage effort.

JIT-Fine (Ni et al., 2022a) introduced the first unified framework to perform both JIT-SDP and JIT-DL. Built on a fine-tuned CodeBERT with an

attention mechanism, it estimates the contribution of each token to commit-level classification and ranks lines by their defect likelihood.

Chen et al. (2024) advanced the field with a dedicated Defect Localization Network (DLN) that explicitly incorporates line-level defect labels for supervised learning, representing a shift from previous indirect approaches.

Huang et al. (2024) proposed JIT-Block, addressing input-length limitations in pretrained models. Using CodeBERT’s attention mechanism, it computes token contributions to commit-level predictions, aggregates them into line-level scores, and ranks lines by defect probability, overcoming truncation issues in large diffs.

Finally, JIT-CF (Ju et al., 2025) unifies JIT-SDP and defect localization using CodeBERT embeddings enhanced through contrastive learning. By amplifying distinctions between defect inducing and non-defective code – even when syntactic differences are subtle – JIT-CF delivers more accurate and robust line-level localization within commits.

Prior JIT-DL approaches rely on attention mechanisms from fine-tuned models or specialized contrastive frameworks. However, these models require supervised training and cannot generalize beyond the project or dataset they were trained on. Furthermore, all such methods depend on the quality and completeness of their training labels.

### *2.5.2 General Software Defect Localization with Large Language Models*

Wu et al. (2023) conducted a large-scale empirical study evaluating ChatGPT 3.5 and ChatGPT 4 for fault localization (FL) in Defects4J (Just et al., 2014), comparing them with SmartFL (Zeng et al., 2022), spectrum-based fault localization (SBFL), and mutation-based fault localization. They demonstrate that ChatGPT 4, augmented with runtime information such as failing tests and error logs, outperforms SmartFL. However, performance degrades sharply as the code context expands from function to class level, highlighting the sensitivity of LLMs to input length. To validate against dataset-specific overfitting, the authors introduce the StuDefects dataset, where ChatGPT 4 continues to maintain its advantage.

Kang et al. (2023) introduce AutoFL, an autonomous FL framework that uses LLMs for iterative code exploration through OpenAI’s function-call API, overcoming prompt-length limitations. AutoFL requires only a single failing test and performs fault localization followed by explanation generation. In Defects4J, it surpasses previous methods and outperforms SBFL under the same single-test constraints, demonstrating the effectiveness of functional augmentation for scalable LLM-driven FL.

Hossain et al. (2024) propose Toggle, a modular token-granular bug localization and repair framework. It employs a CodeT5-based encoder for precise localization, an optional adjustment model to resolve tokenizer mismatches, and a generative LLM for bug repair. Experiments on CodeXGLUE (Lu et al., 2021) and Defects4J show state-of-the-art Top-k performance, confirming the

benefits of token-level reasoning, modularity, and the incorporation of contextual cues (e.g., review comments, defect-inducing line hints).

Subramanian (2024) present BugLLM, a zero-shot bug localization method that combines LLM with embedding-based semantic search. After Abstract Syntax Tree (AST)-driven segmentation and vector indexing of a project’s codebase, one LLM converts bug reports into technical queries, while another verifies retrieved candidates through reasoning. Evaluated on six large Java systems, BugLLM achieves competitive Top-5 accuracy, often surpassing classic IR methods, and offers strong explainability through chain-of-thought prompting, enhancing trust and interpretability in bug localization.

Although recent studies apply LLMs to general fault localization, their evaluations are almost exclusively quantitative, focusing on Top-k metrics or coarse-grained localization (file or method level). Critically, none of these works investigates the underlying reasoning failures of LLMs

### 3 Methodology

#### 3.1 Dataset

For our experiments, we rely on the JIT-Defects4J dataset (Ni et al., 2022a), which has also been adopted in several recent studies (Liu et al., 2024; Chen et al., 2024; Huang et al., 2024; Ju et al., 2025) . This resource provides commit-level information, including messages, source code modifications, and the 14 commit level metrics originally introduced by Kamei et al. (2013). It comprises 21 Java-based open-source projects hosted on GitHub and was specifically curated to disentangle bug-fixing commits that typically combine heterogeneous types of modifications, a known source of noise for defect prediction models. JIT-Defects4J extends the LLTC4J corpus (Herbold and et al., 2020) by incorporating both defect inducing and clean commits as well as line-level annotations, whereas LLTC4J concentrated exclusively on bug-fixing instances. Descriptive statistics of the resulting dataset are presented in Table 1.

For RQ1, the dataset was refined by chronologically ordering the commits and estimating the introduction times of the bugs through the first associated fixing commit, obtained from CommitGuru, as explicit detection timestamps were not available. In the continual learning process, some commits initially marked as non-defective were later reclassified as defect-inducing once their corresponding fixes appeared in subsequent training iterations. Both dataset manipulations were necessary to emulate the real-world arrival of commits over time.

Furthermore, in the original dataset, we encountered an issue that caused random behavior in the predictions, due to the fact that the added/removed lines were stored using the serialized `set` data structure from the Python language. As sets do not have an intrinsic order, they cause different predictions each time the same model is run. We fixed this issue by converting the sets to lists.

Table 1: Statistics of studied datasets: JIT-Defects4J. Adapted from Ni et al. (2022a).

Project	No. of Defect Inducing Commits	No. of Clean Commits	% Ratio (Bugs/ALL)
ant-ivy	332	1,439	18.75%
commons-bcel	60	765	7.27%
commons-beanutils	37	574	6.06%
commons-codec	36	725	4.73%
commons-collections	50	1,773	2.74%
commons-compress	178	1,452	10.92%
commons-configuration	155	1,683	8.43%
commons-dbcp	58	979	5.59%
commons-digester	19	1,060	1.76%
commons-io	73	1,069	6.39%
commons-jcs	88	743	10.59%
commons-lang	146	2,823	4.92%
commons-math	335	3,691	8.32%
commons-net	117	1,004	10.44%
commons-scxml	47	497	8.64%
commons-validator	36	562	6.02%
commons-vfs	114	996	10.27%
giraph	163	681	19.31%
gora	39	514	7.05%
opennlp	91	995	8.38%
parquet-mr	158	962	14.11%
<b>ALL</b>	<b>2,332</b>	<b>24,987</b>	<b>8.54%</b>

### 3.2 Evaluation metrics

**RQ1 How do pre-trained language models perform in comparison to traditional machine learning approaches for continual within-project and cross-project Just-in-Time Software Defect Prediction (JIT-SDP)?**

For online defect prediction, we use the *G-mean* (geometric mean of recall for positive and negative classes) and the absolute difference between class recalls  $|R_1 - R_0|$ , both commonly used in online JIT-SDP studies (Cabral et al., 2019, 2023; Cabral and Minku, 2023; Tabassum et al., 2023; Song et al., 2023; Wu et al., 2025; Liu et al., 2025).

**RQ2 How do Large Language Models (LLMs) compare with attention-based mechanisms for Just-in-Time Defect Localization (JIT-DL)?**

For defect localization, we mainly use the following metrics, borrowed from previous work (Ni et al., 2022a; Huang et al., 2024; Ju et al., 2025; Chen et al., 2024):

- *Recall@20%*: The proportion of real defect-inducing lines found in the top 20% items of the ranked list produced by the localization technique.

- *Effort@20%*: The number of clean lines to be analyzed before encountering the top 20% real defect inducing lines.
- *IFA (Initial false alarms)*: The number of non-faulty lines that appear before the first true faulty line in the ranked list produced by the localization technique.

Further, we also use Top-5 and Top-10 (Ni et al., 2022a; Huang et al., 2024; Ju et al., 2025; Chen et al., 2024), however, in a different way as calculated in the baselines considered<sup>1</sup>

The Top-N (or Top-k) metric serves as a pivotal evaluation criterion, widely used in software engineering research, particularly for assessing the performance of techniques aimed at Fault Localization and Just-in-Time Defect Localization (Chen et al., 2024; Liang et al., 2022). The primary function of this metric is to quantify the effectiveness of an approach by determining the proportion of cases where the actual faulty entity — be it a code line, statement, or file — is successfully identified within the candidates with the highest ranking  $N$  suggested by the diagnostic model (Liang et al., 2022; Niu et al., 2024).

Specifically, the calculation, known as Top-k Accuracy, measures the percentage of defect-inducing changes for which at least one actual faulty element is ranked within the Top- $k$  positions of the returned list. For each instance examined, the metric yields a binary result: 1 if the fault is found within the threshold  $k$ , and 0 otherwise, with the overall accuracy representing the average across the dataset. This focus on low ranks directly addresses practical constraints, as developers typically dedicate only a limited amount of time and effort to inspecting a set of prioritized statements, making thresholds like Top 1, Top 5, and Top 10 highly relevant and commonly reported.

### 3.3 Baseline Classifiers

**RQ1 How do pre-trained language models perform in comparison to traditional machine learning approaches for continual within-project and cross-project Just-in-Time Software Defect Prediction (JIT-SDP)?**

As a baseline for incremental defect prediction, we use Batch Oversampling Rate Boosting (BORB) (Cabral et al., 2023). This method is derived from the online Oversampling Rate Boosting (ORB) algorithm (Cabral et al., 2019), while preserving its central idea of adaptively tuning resampling rates in response to the evolving class imbalance. ORB applies this mechanism in a fully

---

<sup>1</sup> When we compared our work with the baselines, we noticed that all works share the same formulae for Top-5 and Top-10. These formulae divide the number of real faulty lines by the total number of lines in the commit. We understand that this calculation differs from the most common one found in the literature. We maintained this calculation in the initial experiments for fair comparison with the baselines, but later revised it to reflect a more standardized and realistic calculation.

online scenario, incrementally updating the model as each instance arrives. In contrast, BORB processes accumulated batches of data and periodically reinitializes training, allowing models to be built anew. This batch-oriented strategy enables multiple re-visits to past data, which can yield gains in predictive accuracy. Importantly, both algorithms incorporate the notion of verification latency by restricting training inputs to information that would have been observable at the time of prediction.

The decision to adopt BORB was motivated by its nature as an incremental learning framework that takes advantage of batch-based training, departing from a strictly online paradigm where learning occurs in a single instance at a time, a configuration unsuitable for deep neural architectures. To our knowledge, BORB is the only existing framework that employs this strategy.

## RQ2 How do Large Language Models compare with attention-based mechanisms for Just-in-Time Defect Localization (JIT-DL)?

In selecting the baselines for the JIT-DL task, we deliberately excluded approaches that were explicitly fine-tuned for line-level defect localization. This exclusion was a methodological decision made to ensure the validity and fairness of the comparative analysis. Because our LLM-based investigation employs a zero-shot, prompt-driven approach without task-specific fine-tuning, comparing it against models optimized for line-level localization would introduce a confounding advantage unrelated to architectural capability. Instead, by benchmarking our approach against attention-based models that are also not fine-tuned for this task, we isolate and assess its intrinsic reasoning and generalization abilities.

On this basis, we compare state-of-the-art LLMs with the following baselines:

*JIT-Fine* (*Ni et al., 2022a*): JIT-Fine is a unified model designed to address both JIT-SDP and JIT-DL simultaneously, accurately locating the defect position at the line level. Crucially, JIT-Fine achieves line-level localization by utilizing the attention mechanism embedded within the fine-tuned CodeBERT model, which calculates the contribution (weights) of each input code token towards the final commit classification, enabling the ranking of suspicious defect inducing lines.

*JIT-Smart<sub>ATTN</sub>* (*Chen et al., 2024*): JIT-Smart<sub>ATTN</sub> denotes the configuration of the JIT-Smart framework that leverages only the CodeBERT attention weights for the location of line-level defects. This strategy parallels that of prior methods, such as JIT-Fine, in which defect inducing code lines are localized post hoc, based on the predictions produced during the JIT-SDP phase. As explained above, JIT-Smart’s Defect Localization Network (DLN) was excluded from our evaluation because its task-specific fine-tuning would compromise the fairness and methodological consistency of the comparison.

Table 2: Summary of Baseline Classifiers

Baseline	Task	Model / Model Architecture
BORB	Incremental JIT-SDP	Multilayer Perceptron (MLP), Logistic Regression (LR), Naïve Bayes (NB), Iterative Random Forest (IRF)
JIT-Fine	JIT-DL	CodeBERT
JIT-SmartATTN	JIT-DL	CodeBERT
CodeT5+	JIT-DL	Encoder component of CodeT5+ encoder-decoder architecture
UniXCoder	JIT-DL	Encoder-only with mask attention matrices and prefix adapters

*CodeT5+:* CodeT5+ was initially chosen as one of the baselines because it was among the top-performing models in our experiments on defect prediction (Monteiro et al., 2025). Although CodeT5+ does not contain an explicit [CLS] token, as in BERT-based language models, we still use the first encoded token as the head of the classification layer. Therefore, we maintain the default practice of inspecting the weights of the first token attention heads.

*UniXCoder:* In the same way as in CodeT5+, UniXCoder was also among the top performers in defect prediction experiments (Monteiro et al., 2025), so we keep the same default strategy of using the first encoded token attention weights.

We also initially considered JIT-Block (Huang et al., 2024) and JIT-CF (Ju et al., 2025). Regarding JIT-Block, its authors reconstructed the dataset (JIT-Defects4J) into the changed block format, which preserves the relative positional information between added and deleted code lines — information lost in traditional datasets — thus facilitating the model’s ability to learn the semantic meaning of code changes. So, as the dataset was changed, it would not be possible to conduct a fair comparison. Finally, according to its published results, JIT-CF does not achieve better results than JIT-Smart.

A consolidated overview of the baseline classifiers is presented in Table 2.

### 3.4 Description of the Experiments

**RQ1 How do pre-trained language models perform in comparison to traditional machine learning approaches for continual within-project and cross-project Just-in-Time Software Defect Prediction (JIT-SDP)?**

CodeFlowLM was developed in a modularized architecture composed of two components:

- A training loop responsible for grouping the commits received in chronological order in training batches. The number of commits in each training batch  $n$  is a configurable parameter. In our experiments, we used  $n = 50$  for a fair comparison with the BORB default training batch size. This component is also responsible for managing a training queue of recently received commits, and a training pool for commits that 1) passed the latency verification period without any defect detection or 2) were found to be defect inducing (e.g., due to the finding of some defect associated with the given commit).
- A base learner, assumed here to be a neural model such as a pretrained language model (e.g., CodeBERT, CodeT5, CodeT5+, UniXCoder, CodeLlama, etc.). In the current implementation, the base learner may even be an external and independent program that is called by the training loop every training interval.

We experimented with CodeT5+ and UniXCoder as base learners in the within-project setting. UniXCoder achieved shorter training times, but CodeT5+ achieved superior classification performance by a large margin, despite its larger size and longer training times. So, we chose CodeT5+ as the base learner for the cross-project setting experiments. We implemented the same oversampling mechanism proposed by Cabral et al. (2019), and used LoRA as the fine-tuning technique.

Given the mismatch in dimensionality – 512 for the semantic embedding and 14 for the expert attributes – we follow Liu et al. (2024) by first lifting the expert features with a feedforward MLP to a higher-dimensional embedding. We then concatenate this transformed vector with the 512-D semantic representation to obtain a 1024-D joint embedding, which serves as input to a final linear classifier.

Each training batch consists of 10 training epochs, with a batch size of 16 and a learning rate of  $10^{-4}$ . The values for these hyperparameters were inherited from previous experiments (Monteiro et al., 2025; Liu et al., 2024). Our experiments were mainly performed on A100 GPUs.

In contrast to BORB, CodeFlowLM employs incremental fine-tuning rather than full retraining from the initial checkpoint, thereby mitigating instability caused by weight reinitialization. Furthermore, CodeFlowLM avoids BORB’s costly hyperparameter optimization by retaining default parameter choices. To ensure comparability, both methods apply a 90-day deferral period before clean commits are considered.

Finally, for the cross-project setting, we fine-tuned a model for each of the target projects evaluated during the experiment. In order to do this, for each training step, we considered the following training data:

- all the target project’s commits since the beginning of the project until the current time step;
- the commits from the other projects that occurred since the last training step and the current time step. Here, the full data since the beginning

of the other projects is not used because it would lead to unacceptable execution times for the experiments.

In the first training step, as we do not have data from the target project, we use only data from the other projects by selecting commits that occurred before the target project’s initial commit. In this way, the model can start to learn an initial fine-tuned checkpoint even without any examples from the target project.

In the same way as in the within-project setting, in each training step, we respect latency verification and consider only clean commits that have passed the latency verification waiting time, as well as defect inducing commits that have already been labeled as defective, taking into account the date of the first fixing commit.

In BORB, the same strategy is employed to train different cross-project models for each target project; therefore, this was our initial choice to enable a fair comparison.

#### RQ2 How do Large Language Models compare with attention-based mechanisms for Just-in-Time Defect Localization (JIT-DL)?

For just-in-time defect localization, we used the following state-of-the-art models: OpenAI GPT-5<sup>2</sup>, Anthropic Claude Sonnet 4.5<sup>3</sup>, and Google Gemini 2.5 Pro (Google, 2025).

First, we collected the CodeT5+ predictions for the JITDefects4J test split and iterated over each commit correctly predicted as positive by the model. For each commit, we used the prompt shown in Listing 1, inspired by Wu et al. (2023).

Listing 1: Prompt for Defect Inducing Diff Analysis

You are a static analysis and fault localization expert.  
The following GitHub diff introduces a bug. Please analyze specifically the changed lines in the diff for potential bugs. DO NOT consider the commented lines.  
Return the results in JSON format, consisting of a single JSON object with two fields: “intentOfThisCommit” (describing the intended purpose of the commit), and “faultLocalization” (an array of JSON objects). The “faultLocalization” array should contain ten JSON objects, each with four fields: “lineNumber” (indicating the line number of the suspicious code), “codeContent” (showing the actual code), “reason” (explaining why this location is identified as potentially faulty) and “score” (a number between 1 and 10 indicating the suspicion level).  
Be concise and deterministic – avoid generic language (“may be wrong”) and point to \*specific failure modes\*. Prioritize statement lines over declaration lines.

As we ask the LLM to assign a defect score to each line, we sort the returned list of fault localizations in descending order of scores.

<sup>2</sup> GPT-5 System Card: <https://cdn.openai.com/gpt-5-system-card.pdf>

<sup>3</sup> Claude Sonnet 4.5 System Card: <https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf>

Finally, we compared the LLMs' responses with JITDefect4J labels for each commit line. Ultimately, we calculated the average values for each metric, considering the entire set of commits in the test split.

For the best-performing LLM, we repeated the experiment four times to account for randomness.

### RQ3 Despite their near-human capabilities, in which situations do Large Language Models still underperform in fault localization?

To answer this research question, we divided the analysis into two scenarios: false positives and false negatives.

Regarding false positives, we consider the answers from the best-performing LLM. For all the commits in the test split, we collected the clean lines classified as defect-inducing by the LLM (false positives). We sorted the set of false positives in decreasing order according to the score given by the LLM. We then analyzed false positives with a maximum score (score = 10). In this step, we group the most common types of errors in the sample. Finally, we manually inspect some examples of each error group to gain further insight.

Finally, in relation to false negatives, we extracted the lines labeled as defect-inducing in the dataset which were not pointed out in the lists returned by the LLM (limited to ten faults per commit according to the prompt shown in Listing 1). We prioritized for manual inspection the longest lines and, again, we grouped them in the most common types of errors.

## 4 Results and Discussion

This section reports on experiments aimed at addressing our research questions, restated here.

### 4.1 Continual Just-in-Time Software Defect Prediction

#### RQ1: How do pre-trained language models perform in comparison to traditional machine learning approaches for continual within-project and cross-project Just-in-Time Software Defect Prediction (JIT-SDP)?

This RQ examines models such as CodeT5+ and UniXCoder in incremental learning settings, comparing them with a state-of-the-art baseline (Cabral et al., 2023) for continual Just-in-Time Defect Prediction, introducing CodeFlowLM as a new framework for incremental JIT-SDP with pre-trained code language models. We investigated both within-project and cross-project performance.

Table 3: Comparison between CodeFlowLM (CodeT5+ and UniCoder) versus BORB (different base learners) in within-project setting (WP)

Project	BORB			CodeFlowLM			
	Model	UniCoder		CodeT5+			
		G-Mean	$ R1 - R0 $	G-Mean	$ R1 - R0 $	G-Mean	$ R1 - R0 $
ant-ivy	LR	0.591	0.272	0.612	0.403	<b>0.680</b>	<b>0.197</b>
commons-bcel	LR	0.467	0.428	0.487	0.356	<b>0.567</b>	<b>0.344</b>
commons-beanutils	MLP	0.397	0.448	0.516	0.316	<b>0.531</b>	<b>0.324</b>
commons-codec	LR	<b>0.552</b>	0.192	0.459	0.532	0.546	<b>0.183</b>
commons-collections	IRF	<b>0.481</b>	<b>0.326</b>	0.154	0.884	0.183	0.798
commons-compress	MLP	<b>0.534</b>	<b>0.265</b>	0.515	0.512	0.533	0.442
commons-configuration	LR	0.580	0.468	0.633	0.435	<b>0.771</b>	<b>0.152</b>
commons-digester	MLP	0.448	0.436	0.591	0.505	<b>0.591</b>	<b>0.162</b>
commons-jcs	MLP	0.442	<b>0.425</b>	0.411	0.579	<b>0.516</b>	0.451
commons-lang	LR	0.555	<b>0.337</b>	0.548	0.582	<b>0.607</b>	0.480
commons-math	MLP	0.569	<b>0.152</b>	0.556	0.547	<b>0.635</b>	0.375
commons-net	IRF	0.472	0.423	0.463	0.504	<b>0.524</b>	<b>0.301</b>
commons-scxml	LR	0.443	0.530	<b>0.540</b>	<b>0.252</b>	0.515	0.298
commons-validator	MLP	0.487	0.397	0.365	<b>0.345</b>	<b>0.511</b>	0.390
commons-vfs	NB	0.481	0.546	0.509	0.379	<b>0.565</b>	<b>0.286</b>
giraph	LR	0.553	0.317	0.545	0.337	<b>0.580</b>	<b>0.242</b>
gora	IRF	0.411	0.370	0.571	<b>0.204</b>	<b>0.572</b>	0.295
opennlp	LR	0.410	<b>0.319</b>	<b>0.418</b>	0.520	0.356	0.591
parquet-mr	NB	0.533	0.415	0.441	0.572	<b>0.564</b>	<b>0.397</b>

#### 4.1.1 Incremental Within-Project JIT-SDP

Table 3 reports the G-Mean and  $|R1 - R0|$  metrics in CodeFlowLM for a selected subset of JITDefects4J projects compared to BORB (Cabral et al., 2023). The calculation of the metrics followed a prequential evaluation scheme with a fading factor of 0.99, as recommended in Cabral et al. (2023). The six BORB base learners were tested on each project; however, for simplicity, only the highest mean G-means results are presented, with full logs to be released in the project code repository. Across the evaluated projects, CodeFlowLM, employing CodeT5+ or UniCoder, achieved the best G-Mean scores in 16 out of 19 cases by at least 5%. CodeT5+ exceeded the baseline in 14 projects, while UniCoder exceeded it in 9, highlighting the advantages of CodeT5+'s larger parameter count (770 M) for incremental multimodal learning over code, textual, and tabular input.

Across the 19 projects, CodeFlowLM achieved higher G-Mean in 16 cases, with gains ranging from 5% to 39% in WP settings. This consolidated pattern strongly suggests that PLMs can be successfully adapted to incremental, latency-aware scenarios, outperforming traditional models even in the presence of concept drift.

#### 4.1.2 Incremental Cross-Project JIT-SDP

Table 4 reports the G-Mean and  $|R1 - R0|$  metrics. CodeFlowLM, using CodeT5+, achieved the highest G-Mean in 18 of 19 projects, with improvements ranging from 10% to 68%. In terms of R1, CodeFlowLM yielded better values in 16 of 19 projects, by at least 5% up to 91%, demonstrating robust

Table 4: Comparison between CodeFlowLM (CodeT5+) versus BORB (different base learners) in cross-project setting (CP)

Project	BORB					CodeFlowLM (CodeT5+)			
	Model	G-Mean	$\ R1 - R0\ $	R1	R0	G-Mean	$\ R1 - R0\ $	R1	R0
ant-ivy	IRF	0.629	0.142	0.637	0.662	<b>0.786</b>	<b>0.102</b>	<b>0.747</b>	<b>0.831</b>
commons-bcel	IRF	0.535	0.270	0.451	0.710	<b>0.671</b>	<b>0.246</b>	<b>0.617</b>	<b>0.763</b>
commons-beanutils	IRF	0.482	0.400	0.402	<b>0.802</b>	<b>0.625</b>	<b>0.266</b>	<b>0.594</b>	0.785
commons-codec	IRF	0.561	<b>0.290</b>	<b>0.478</b>	0.766	<b>0.626</b>	0.483	0.427	<b>0.917</b>
commons-collections	IRF	<b>0.460</b>	0.345	<b>0.387</b>	0.711	0.414	0.618	0.247	<b>0.865</b>
commons-compress	MLP	0.593	<b>0.160</b>	0.549	0.686	<b>0.683</b>	0.318	<b>0.578</b>	<b>0.848</b>
commons-configuration	IRF	0.576	<b>0.212</b>	0.637	0.574	<b>0.785</b>	0.270	<b>0.684</b>	<b>0.915</b>
commons-digester	MLP	0.349	<b>0.383</b>	0.282	0.591	<b>0.586</b>	0.450	<b>0.439</b>	0.889
commons-jcs	LR	0.675	<b>0.182</b>	0.654	0.768	<b>0.749</b>	0.224	<b>0.713</b>	<b>0.807</b>
commons-lang	MLP	0.524	<b>0.217</b>	0.461	0.676	<b>0.676</b>	0.347	<b>0.545</b>	<b>0.882</b>
commons-math	MLP	0.603	<b>0.132</b>	<b>0.608</b>	0.653	<b>0.663</b>	0.379	0.508	<b>0.887</b>
commons-net	IRF	0.493	<b>0.360</b>	0.367	0.725	<b>0.614</b>	0.423	<b>0.472</b>	<b>0.853</b>
commons-saxml	MLP	0.580	0.230	0.568	0.719	<b>0.693</b>	<b>0.151</b>	<b>0.638</b>	<b>0.764</b>
commons-validator	MLP	0.450	<b>0.343</b>	0.377	<b>0.707</b>	<b>0.678</b>	0.348	<b>0.719</b>	0.694
commons-vfs	IRF	0.557	0.337	0.450	0.787	<b>0.705</b>	<b>0.212</b>	<b>0.626</b>	<b>0.808</b>
giraph	IRF	0.657	0.160	0.654	0.731	<b>0.778</b>	<b>0.097</b>	<b>0.819</b>	<b>0.749</b>
gora	MLP	0.562	0.240	0.543	<b>0.715</b>	<b>0.724</b>	<b>0.064</b>	<b>0.744</b>	<b>0.715</b>
opennlp	IRF	0.534	<b>0.168</b>	0.490	0.637	<b>0.673</b>	0.364	<b>0.598</b>	<b>0.806</b>
parquet-mr	LR	0.622	0.257	0.566	0.756	<b>0.723</b>	<b>0.163</b>	<b>0.653</b>	<b>0.806</b>

classification performance for the positive class. For R0, CodeFlowLM also achieved better values in 16 projects, by up to 59%, demonstrating that the model is capable of achieving high recall in both classes. We observed that BORB achieves a smaller difference between positive and negative classes in most projects (11 out of 19) compared to CodeFlowLM. However, the greater difference between R1 and R0 in CodeFlowLM does not result in the sacrifice of any of the classes, as can be seen in both recall metrics.

G-mean declined in just one case – commons-collections – which suffered from an acute early-phase imbalance (only a single positive instance in the first 300 examples). By contrast, commons-digester, despite exhibiting the strongest overall imbalance, delivered the largest G-mean gain over BORB, indicating that the short-horizon imbalance at the start of commons-collections – not the aggregate skew – is the primary factor delaying learning there. In CodeFlowLM, models are incrementally fine-tuned rather than reinitialized; therefore, exposure to initially labeled negative examples may leave a residual influence even if those examples are later relabeled as positive. BORB, in turn, rebuilds its models from scratch. However, CodeFlowLM’s consistently superior G-mean on most projects implies that any such effect is minimal in practice.

*[RQ1]: Pre-trained language models, particularly CodeFlowLM with CodeT5+ and UnixCoder, consistently outperform traditional machine learning approaches such as BORB in continual Just-in-Time Software Defect Prediction. In the within-project setting, CodeFlowLM achieved the highest G-Mean in 16 of 19 projects, demonstrating superior adaptability and stability under incremental learning. In the cross-project setting, it outperformed BORB in 18 of 19 projects, with G-Mean gains of 10–68% and balanced recalls for both positive and negative classes. Despite slightly higher  $|R_1 - R_0|$  differences, CodeFlowLM maintained strong recall in both classes without sacrificing performance, confirming that pre-trained language models deliver higher generalization and robustness for continual within- and cross-project JIT-SDP tasks compared to traditional incremental learners. These findings demonstrate that continual fine-tuning of PLMs is not only feasible but consistently superior to traditional machine-learning-based incremental strategies, providing the first empirical evidence that PLMs can operate effectively under realistic JIT-SDP constraints.*

## 4.2 Fault Localization

### 4.2.1 Quantitative Analysis

#### RQ2 How do Large Language Models compare with attention-based mechanisms for Just-in-Time Defect Localization (JIT-DL)?

This research question addresses the problem of fault localization (FL) specifically in the context of just-in-time defect detection. Recent solutions are based on common techniques such as 1) attention-based fault localization (Ni et al., 2022a; Huang et al., 2024; Ju et al., 2025) or 2) multitask learning in both classifying commits in clean/defect-inducing and pointing out the possible bug locations in tokens or lines (Chen et al., 2024). We intend to investigate the capabilities of Large Language Models (LLMs) in this task from a quantitative perspective, as these models have recently been successfully applied in broader contexts, including file- and method-based software defect prediction and Automatic Program Repair (APR).

According to Table 5<sup>4</sup>, the LLM with the highest performance – GPT-5 – is competitive with the JIT-SmartATTN baseline in Recall@20% and Effort@20% metrics. We repeated the experiment with JIT-SmartATTN four times to take into account different random seeds.

Surprisingly, despite showing the best performance, JIT-SmartATTN demonstrates more instability when trained with different seeds (i.e., network initialization), as indicated by the standard deviation values. However, GPT-5,

<sup>4</sup> Due to time and cost constraints, the number of experiments' repetition of the different models may vary.

Table 5: Comparative results of JIT-DL models

Model	Top-5 ↑	Top-10 ↑	Recall@20% ↑	Effort@20% ↓	IFA ↓
JIT-Smart ATTN	<b>0.691</b> ( $\sigma \pm 0.03$ )	<b>0.811</b> ( $\sigma \pm 0.03$ )	<b>0.337</b> ( $\sigma \pm 0.05$ )	0.248 ( $\sigma \pm 0.02$ )	<b>5.360</b> ( $\sigma \pm 3.00$ )
GPT-5	0.616 ( $\sigma \pm 0.01$ )	0.759 ( $\sigma \pm 0.02$ )	0.318 ( $\sigma \pm 0.01$ )	<b>0.234</b> ( $\sigma \pm 0.01$ )	10.711 ( $\sigma \pm 0.92$ )
Gemini 2.5 Pro	0.543	0.683	0.291	0.251	12.317
UniXCoder	0.522 ( $\sigma \pm 0.01$ )	0.657 ( $\sigma \pm 0.01$ )	0.278 ( $\sigma \pm 0.01$ )	0.271 ( $\sigma \approx 0$ )	13.280 ( $\sigma \pm 0.36$ )
Claude Sonnet 4.5	0.529	0.701	0.273	0.265	12.520
JITFine	0.511 ( $\sigma \pm 0.03$ )	0.683 ( $\sigma \pm 0.02$ )	0.217 ( $\sigma \pm 0.02$ )	0.322 ( $\sigma \pm 0.02$ )	12.185 ( $\sigma \pm 1.21$ )
CodeT5+	0.465 ( $\sigma \pm 0.04$ )	0.631 ( $\sigma \pm 0.04$ )	0.150 ( $\sigma \pm 0.04$ )	0.388 ( $\sigma \pm 0.05$ )	15.048 ( $\sigma \pm 2.43$ )

Table 6: Fine-tuning of Small to Medium-Sized LMs (summary from Monteiro et al. (2025))

Model	ROC-AUC	F1
UniXCoder	<b>0.901</b>	<b>0.519</b>
CodeT5+	0.899	<b>0.519</b>
JIT-Smart	0.885	0.486

when called with the default temperature setting, exhibits a more stable performance.

Conversely, a previous work (Monteiro et al., 2025) demonstrates that, despite a better accuracy in defect localization, JIT-Smart shows inferior F1 and ROC-AUC compared to UniXCoder and CodeT5+ in the JITDefects4J dataset. The results are shown in Table 6. In addition, the same authors demonstrated that both CodeT5+ and UniXCoder outperform JIT-Smart in cross-project JIT-SDP on the same dataset.

Finally, the significantly worse results of LLMs in metrics such as Top-5, Top-10, and IFA motivate the investigation carried out in the next research question.

[RQ2]: Unlike PLM-based models, LLMs do not rely on supervised fine-tuning for defect localization; instead, their performance reflects raw reasoning capabilities. This makes their comparison with attention-based methods particularly revealing. LLMs such as GPT-5 demonstrate competitive performance with attention-based models like JIT-SmartATTN in key localization metrics (Recall@20% and Effort@20%), while exhibiting greater stability across random runs. However, JIT-SmartATTN achieves higher accuracy but exhibits instability across different training random seeds and yields inferior results in broader evaluation metrics (F1, ROC-AUC) compared to pre-trained code models, such as CodeT5+ and UniXCoder. Overall, LLMs demonstrate promising robustness and comparable defect localization capabilities without fine-tuning, but still lag behind specialized models in ranking precision metrics (Top-5, Top-10, IFA). While LLMs show stronger robustness and competitive Recall@20%, their weaker Top-k and IFA results suggest that they prioritize semantic risk over syntactic precision, a pattern further analyzed in RQ3.

#### 4.2.2 Qualitative Analysis

##### RQ3 Despite their near-human capabilities, in which situations do Large Language Models still underperform in fault localization?

In this RQ, we perform an error analysis to identify the main classes of errors that LLMs still incur in defect localization.

*False Positives Analysis* We initially found a total of 1,587 false positives, 92 of which had a maximum score = 10. As described in Section 3.4, we used that subsample for our initial investigation.

The preliminary analysis reveals common types of mistakes, as shown in Table 7.

Table 7: False Positive Categories in JIT-DL Classification

Category	Short Description	Estimated %
Apparent compilation errors	Valid API usage flagged as invalid.	17
Unchecked null access	Potential null dereference but logically safe.	14
Binary incompatibilities	Interface change misclassified as error.	11
Reflection/type-related errors	Intended reflective or cast operations.	10
Non-observable logic conditions	The model assumes that a logic expression is incorrect, ignoring semantic context.	10
Potential arithmetic errors	Division/overflow flagged but constrained.	8
Deliberate configuration choices	Intentional constants for compatibility.	8
Test IO and path issues	False positives from mock/test paths.	7

Based on preliminary observation, the following potentially detrimental hypotheses will be further analyzed.

- H1 (human bias): LLMs tend to be over-conservative when code reviewing, like a human software developer who tends to label a large number of code lines as suspect.
- H2 (lack of enough context): LLMs may lack extra context that otherwise would warrant correctness (e.g., correct variable initialization).
- H3 (dataset mislabeling): Some bugs might not have appeared at the time the dataset was originally built, but may have appeared later. Therefore, some execution flow may not have been run, potentially causing a bug to go undetected. The hypothesis is particularly important even for fine-tuned models because of the risk of overfitting to an outdated dataset.

For each of the above hypotheses, typical code snippets extracted from the JIT-Defects4J dataset are used to depict the scenario.

*H1 (human bias)* According to this hypothesis, some errors may occur primarily because the LLM acts like a human being who is unaware of the execution configuration or the expected execution flow. Some concrete scenarios range from potential arithmetic errors to type-related errors. In cases involving potential arithmetic errors, the LLM labeled as defect-inducing lines that perform operations such as dividing by a variable that could be zero, even though the code logic may guarantee that this situation never occurs. (Listing 2). In reflection/type-related errors, the LLM assumes that variable castings are potentially risky, even when they occur in controlled situations. For example, in Listing 3, a potential casting error is reported, but at the current version of the ScriptObjectMirror class, it implements the Bindings interface<sup>5</sup>.

*H2 (lack of enough context)* Some errors may occur primarily due to a lack of context, such as false positives related to variable initialization and concerns about API versions. For example, variables pointed as potentially null might have been correctly initialized before that point in the code. Regarding concerns about API versions, constructors and method overloading may be available in older or more recent library versions, but the LLM does not have access to this information in the standard prompt format based on diffs. This can lead, for example, to errors in the “Apparent compilation errors” and “Binary incompatibilities” categories.

*H3 (dataset mislabeling)* The following errors may occur primarily due to poor dataset labeling because of error conditions that had not been observed at the time the dataset was labeled, such as, for example, a non-observable logic condition: the next listing shows an example of an execution flow that may not have been run, potentially causing a bug to go undetected. In the example shown in Listing 7, clearly, if arrayObject is equal to null, the call to arrayObject.getClass()... will be tried, and a NullPointerException will be thrown.

---

<sup>5</sup> ScriptObjectMirror class documentation: <https://docs.oracle.com/javase/8/docs/jdk/api/nashorn/jdk/nashorn/api/scripting/ScriptObjectMirror.html>

Listing 2: Examples of “Potential arithmetic errors” error type

```
{
  "lineNumber": 327,
  "codeContent": "sum += FastMath.pow(distA / distB, 2.0 / (
    fuzzyness - 1.0));",
  "reason": "Division by zero and NaN risk: if distB == 0 (point
    coincides with some cluster center), distA/distB is Infinity
    (or 0/0 -> NaN when ...",
  "project": "commons-math"
}

{
  "lineNumber": 159,
  "codeContent": "if (numUpdates % statsPrintFrequency == 0) {",
  "reason": "statsPrintFrequency can be 0 from configuration,
    causing ArithmeticException (modulo by zero). Previously
    this was a constant 10 and safe.",
  "project": "giraph"
}
```

Listing 3: Examples of “Reflection/type-related errors” error type

```
{
  "lineNumber": 321,
  "codeContent": "return (Bindings) bindings.get(NASHORN_GLOBAL);"
  ,
  "reason": "Unsafe cast: in Nashorn, 'nashorn.global' typically
    holds a ScriptObject/ScriptObjectMirror, not a javax.script.
    Bindings. This cast can trigger ClassCastException when
    accessed.",
  "project": "commons-scxml"
}
(...)

{
  "lineNumber": 186,
  "codeContent": "return (Class<T>) getClassMethod.invoke(schema);
  ",
  "reason": "Reflection is invoked on the wrong receiver. The
    method getClass(Schema) belongs to the model instance, but
    this calls it on the Schema object, causing
    IllegalArgumentException at runtime and preventing correct
    datum class detection.",
  "project": "parquet-mr"
}
```

*False Negatives Analysis* Regarding the second part of our investigation, GPT-5 generated a total of 844 false negatives, nearly half the number of false positives, supporting the intuition that GPT-5 favors a more conservative and bug-seeking approach (H1 in the false positives analysis).

Listing 4: Example of “Unchecked null access” error type

```
{
  "lineNumber": 132,
  "codeContent": "Map parentPomProps =
    PomModuleDescriptorBuilder.extractPomProperties(
      parentDescr.getExtraInfo());",
  "reason": "Dereferences parentDescr without verifying it is non-
    null. If the parent could not be resolved, this line will
    throw a NullPointerException.",
  "project": "ant-ivy"
}
```

Listing 5: Example of “Apparent compilation errors” error type

```
{
  "lineNumber": 436,
  "codeContent": "return new LZMACompressorInputStream(in,
    lzmaMemoryLimitKb);",
  "reason": "Unconditionally uses the new two-argument LZMA
    constructor after only checking LZMAUtils.
    isLZMACompressionAvailable(). That check verifies class
    presence but not constructor availability. With older org.
    tukaani.xz versions lacking LZMAInputStream(InputStream,int)
    , this call can trigger NoSuchMethodError at runtime rather
    than a clean CompressorException.",
  "project": "commons-compress"
}
```

Listing 6: Example of “Binary incompatibilities” error type

```
{
  "lineNumber": 484,
  "codeContent": "return compressorStreamProvider.
    createCompressorInputStream(name, in,
    actualDecompressConcatenated);",
  "reason": "Calling the new 3-arg method on providers is binary-
    incompatible with existing third-party
    CompressorStreamProvider implementations compiled against
    the old 2-arg interface, causing AbstractMethodError/
    IncompatibleClassChangeError at runtime when discovered via
    service loading.",
  "project": "commons-compress"
}
```

Listing 7: Example of “Non-observable logic condition” error type

```
{
  "lineNumber": 121,
  "codeContent": "if (arrayObject != null && arrayObject
    instanceof Iterable || arrayObject.getClass().isArray()) {",
  "reason": "Operator precedence bug: when arrayObject is null,
    the right side arrayObject.getClass().isArray() is still
    evaluated, causing a NullPointerException. Should be:
    arrayObject != null && (arrayObject instanceof Iterable ||
    arrayObject.getClass().isArray())",
  "project": "commons-scxml"
}
```

Table 8: Summary of False Negative Categories (Descriptions)

Category	Description	Freq. (%)
I/O operations without validation	Reading or writing data without checking return values or error conditions, risking partial reads or unnoticed failures.	14
Configuration or metadata merging	Combining settings or metadata from multiple sources without proper validation or ordering, creating subtle inconsistencies.	12
Logging hides underlying issues	Logging warnings or errors instead of handling exceptional situations masks real problems in the system.	6

Table 9: Representative Examples for Each False Negative Category

Category	Representative Example (from dataset)
I/O operations without validation	nbrBytesCopied += in.read(b, off + nbrBytesCopied, len - ...);
Configuration or metadata merging	mergeConfigurations(parent.getModuleRevisionId(), parent.getConfigurations());
Logging hiding underlying issues	LOG.warn("Ignoring statistics because created.by...");

A preliminary analysis reveals common types of mistakes, as shown in Tables 8 and 9. Here, we opt to show the most intuitive ones instead of the most common ones. In fact, there are a larger number of false negatives in code lines that contain cast operations (potentially unsafe), numeric computation, index-based control flow in loops, etc. However, the goal of this analysis is to gather information on better prompt instructions. If we incorporate the full spectrum of false-negative categories into the prompt instructions, we risk generating yet more false positives.

*[RQ3]: Despite their strong performance, the qualitative analysis shows that LLMs still underperform in three main situations: conservative human-like bias, where models over-flag lines as suspicious, leading to false positives such as theoretical arithmetic errors, overly cautious reflection/casting warnings, or misinterpreted logic conditions; insufficient contextual information, causing models to misjudge code that is actually safe or compatible due to missing details about variable initialization, API/version differences, or execution flow unavailable in diff-based prompts; and dataset noise or outdated labeling, which creates mismatches between ground truth and actual defect behavior. Furthermore, the false negatives analysis reveals that LLMs may struggle in areas such as unvalidated I/O operations, fragile configuration/metadata merging, or logging patterns that conceal failures. Overall, these findings suggest that LLMs remain susceptible to context gaps and conservative reasoning patterns, highlighting opportunities for enhanced prompt design, context engineering, and refined dataset curation in the evaluation phase to mitigate such errors. This qualitative taxonomy addresses a critical gap in the literature by offering the first systematic analysis of LLM reasoning failures in JIT-DL, providing actionable insights for prompt design, model evaluation, and dataset refinement.*

## 5 Conclusions

This work introduced CodeFlowLM, an incremental learning framework for Just-In-Time Software Defect Prediction (JIT-SDP) built upon pre-trained language models. Across within-project and cross-project experiments, CodeFlowLM — particularly when using CodeT5+ — consistently outperformed traditional incremental learners such as BORB in both predictive accuracy and stability, achieving superior G-Mean values in most evaluated projects. Our findings confirm, for the first time, that PLMs can be incrementally adapted under real JIT-SDP constraints—drift, verification latency, and imbalance evolution—without model restarts or costly hyperparameter searches.

We also extended our investigation to the Just-in-Time Defect Localization (JIT-DL) task, analyzing the performance and limitations of Large Language Models (LLMs) in this context. Among the evaluated models, GPT-5 achieved competitive results compared to attention-based approaches such as JIT-SmartATTN, particularly in the Recall@20% and Effort@20% metrics, while exhibiting greater stability between runs. Nevertheless, attention-based models still outperformed LLMs in ranking-oriented measures such as Top-5 and IFA, suggesting that although LLMs show strong potential for defect localization, further progress is needed to enhance their fine-grained line prioritization capabilities.

Critically, our qualitative investigation fills a key gap in the literature: understanding why LLMs fail. Our analysis revealed that false positives occur

at a higher frequency than false negatives. The false positives produced by LLMs are often caused by (1) human-like conservative bias, leading to over-identification of suspicious lines; (2) insufficient contextual information due to limited diff-based prompts; and (3) dataset labeling noise, where some “clean” commits in JIT-Defects4J should, in fact, be labeled as defect inducing. These findings highlight both the potential and the current limitations of generative models in reasoning about software changes.

Future work will explore drift-adaptive thresholds, multimodal fusion for joint prediction/localization, and an automated audit of JIT-Defects4J labels. Together, these steps move us toward unified, robust, and trustworthy JIT defect analysis pipelines.

## Declarations

**Funding:** Not applicable.

**Ethical Approval** Not applicable.

**Informed Consent** Not applicable.

**Author Contributions:** Monique Louise Monteiro formulated research questions and hypotheses and conducted the experiments. George G. Cabral and Adriano L.I. Oliveira contributed to the formulation of research questions and hypotheses, as well as the discussion of results, and reviewed the manuscript.

**Data Availability Statement:** The reproduction package, including source code and data used in the experiments, is available at [https://github.com/monilouise/codelflowlm\\_jitdl](https://github.com/monilouise/codelflowlm_jitdl).

**Conflicts of Interest** The authors declared that they have no conflict of interest.

**Clinical Trial Number** Not applicable.

## References

- Abu Talib M, Bou Nassif A, Azzeh M, Alesh Y, Afadar Y (2024) Parameter-efficient fine-tuning of pre-trained code models for just-in-time defect prediction. *Neural Computing and Applications* 36(27):16911–16940, DOI 10.1007/s00521-024-09930-5, ISBN: 1433-3058
- Cabral GG, Minku LL (2023) Towards Reliable Online Just-in-Time Software Defect Prediction. *IEEE Transactions on Software Engineering* 49(3):1342–1358, DOI 10.1109/TSE.2022.3175789
- Cabral GG, Minku LL, Shihab E, Mujahid S (2019) Class Imbalance Evolution and Verification Latency in Just-in-Time Software Defect Prediction. In: 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), pp 666–676, DOI 10.1109/ICSE.2019.00076
- Cabral GG, Minku LL, Oliveira AL, Pessoa DA, Tabassum S (2023) An investigation of online and offline learning models for online Just-in-Time

- Software Defect Prediction. *Empirical Softw Engg* 28(5), DOI 10.1007/s10664-023-10335-6, URL <https://doi.org/10.1007/s10664-023-10335-6>, place: USA Publisher: Kluwer Academic Publishers
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. *J Artif Int Res* 16(1):321–357
- Chen X, Xu F, Huang Y, Zhang N, Zheng Z (2024) JIT-Smart: A Multi-task Learning Framework for Just-in-Time Defect Prediction and Localization. *Proc ACM Softw Eng 1(FSE)*, DOI 10.1145/3643727, URL <https://doi.org/10.1145/3643727>
- Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: North American Chapter of the Association for Computational Linguistics, URL <https://api.semanticscholar.org/CorpusID:52967399>
- Feng Z, Guo D, Tang D, Duan N, Feng X, Gong M, Shou L, Qin B, Liu T, Jiang D, Zhou M (2020) CodeBERT: A Pre-Trained Model for Programming and Natural Languages. In: Cohn T, He Y, Liu Y (eds) Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, pp 1536–1547, DOI 10.18653/v1/2020.findings-emnlp.139, URL <https://aclanthology.org/2020.findings-emnlp.139>
- Google (2025) Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. URL <https://arxiv.org/abs/2507.06261>, 2507.06261
- Guo D, Ren S, Lu S, Feng Z, Tang D, Liu S, Zhou L, Duan N, Svyatkovskiy A, Fu S, Tufano M, Deng SK, Clement C, Drain D, Sundaresan N, Yin J, Jiang D, Zhou M (2021) GraphCodeBERT: Pre-training Code Representations with Data Flow. URL <https://arxiv.org/abs/2009.08366>, 2009.08366
- Guo D, Lu S, Duan N, Wang Y, Zhou M, Yin J (2022) UniXcoder: Unified Cross-Modal Pre-training for Code Representation. URL <https://arxiv.org/abs/2203.03850>, 2203.03850
- Guo Y, Gao X, Zhang Z, Chan W, Jiang B (2023) A study on the impact of pre-trained model on Just-In-Time defect prediction. In: 2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security (QRS), pp 105–116, DOI 10.1109/QRS60937.2023.00020
- Guo Y, Gao X, Jiang B (2024) An Empirical Study on JIT Defect Prediction Based on BERT-style Model. URL <https://arxiv.org/abs/2403.11158>, 2403.11158
- Herbold S, et al (2020) A fine-grained data set and analysis of tangling in bug fixing commits. *Empirical Software Engineering* 27, URL <https://doi.org/10.1007/s10664-021-10083-5>
- Hoang T, Khanh Dam H, Kamei Y, Lo D, Ubayashi N (2019) DeepJIT: An End-to-End Deep Learning Framework for Just-in-Time Defect Prediction. In: 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR), pp 34–45, DOI 10.1109/MSR.2019.00016

- Hoang T, Kang HJ, Lo D, Lawall J (2020) CC2Vec: distributed representations of code changes. In: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering, Association for Computing Machinery, New York, NY, USA, ICSE '20, p 518–529, DOI 10.1145/3377811.3380361, URL <https://doi.org/10.1145/3377811.3380361>
- Hossain SB, Jiang N, Zhou Q, Li X, Chiang WH, Lyu Y, Nguyen H, Tripp O (2024) A deep dive into large language models for automated bug localization and repair. Proc ACM Softw Eng 1(FSE), DOI 10.1145/3660773, URL <https://doi.org/10.1145/3660773>
- Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, Attariyan M, Gelly S (2019) Parameter-Efficient Transfer Learning for NLP. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol 97, pp 2790–2799
- Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2022) LoRA: Low-Rank Adaptation of Large Language Models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net
- Huang T, Yu HQ, Fan GS, Huang ZJ, Wu CY (2024) A code change-oriented approach to just-in-time defect prediction with multiple input semantic fusion. Expert Systems 41(12):e13702, DOI <https://doi.org/10.1111/exsy.13702>, URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13702>, <https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.13702>
- Ju X, Cao Y, Chen X, Gong L, Chakma V, Zhou X (2025) Jit-cf: Integrating contrastive learning with feature fusion for enhanced just-in-time defect prediction. Information and Software Technology 182:107706, DOI <https://doi.org/10.1016/j.infsof.2025.107706>, URL <https://www.sciencedirect.com/science/article/pii/S095058492500045X>
- Just R, Jalali D, Ernst MD (2014) Defects4j: a database of existing faults to enable controlled testing studies for java programs. In: Proceedings of the 2014 International Symposium on Software Testing and Analysis, Association for Computing Machinery, New York, NY, USA, ISSTA 2014, p 437–440, DOI 10.1145/2610384.2628055, URL <https://doi.org/10.1145/2610384.2628055>
- Kamei Y, Shihab E, Adams B, Hassan AE, Mockus A, Sinha A, Ubayashi N (2013) A large-scale empirical study of just-in-time quality assurance. IEEE Transactions on Software Engineering 39(6):757–773, DOI 10.1109/TSE.2012.70
- Kang S, An G, Yoo S (2023) A preliminary evaluation of llm-based fault localization. DOI 10.48550/arXiv.2308.05487
- Kim S, Shivaji S, Whitehead J (2025) A Reflection on Change Classification in the Era of Large Language Models. IEEE Transactions on Software Engineering 51(03):864–869, DOI 10.1109/TSE.2025.3539566, URL <https://doi.ieee.org/10.1109/TSE.2025.3539566>,

- place: Los Alamitos, CA, USA Publisher: IEEE Computer Society
- Li Z, Du Q, Zhang H, Jing XY, Wu F (2024) An empirical study of data sampling techniques for just-in-time software defect prediction. *Automated Software Engineering* 31(2):56, DOI 10.1007/s10515-024-00455-8, URL <https://doi.org/10.1007/s10515-024-00455-8>
- Liang H, Hang D, Li X (2022) Modeling function-level interactions for file-level bug localization. *Empirical Software Engineering* 27(7):186, DOI 10.1007/s10664-022-10237-z, URL <https://doi.org/10.1007/s10664-022-10237-z>
- Lin B, Wang S, Liu Z, Liu Y, Xia X, Mao X (2023) CCT5: A Code-Change-Oriented Pre-trained Model. In: Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Association for Computing Machinery, New York, NY, USA, ESEC/FSE 2023, p 1509–1521, DOI 10.1145/3611643.3616339, URL <https://doi.org/10.1145/3611643.3616339>
- Liu S, Keung J, Yang Z, Liu F, Zhou Q, Liao Y (2024) Delving into Parameter-Efficient Fine-Tuning in Code Change Learning: An Empirical Study. In: 2024 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), pp 465–476, DOI 10.1109/SANER60148.2024.00055
- Liu X, Zhou Y, Tang Y, Qian J, Zhou Y (2025) Human-in-the-loop online just-in-time software defect prediction: What have we achieved and what do we still miss? *Sci Comput Program* 244(C), DOI 10.1016/j.scico.2025.103296, URL <https://doi.org/10.1016/j.scico.2025.103296>
- Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. URL <https://arxiv.org/abs/1907.11692>, 1907.11692
- Lu S, Guo D, Ren S, Huang J, Svyatkovskiy A, Blanco A, Clement C, Drain D, Jiang D, Tang D, Li G, Zhou L, Shou L, Zhou L, Tufano M, Gong M, Zhou M, Duan N, Sundaresan N, Deng SK, Fu S, Liu S (2021) Codexglue: A machine learning benchmark dataset for code understanding and generation. URL <https://arxiv.org/abs/2102.04664>, 2102.04664
- Monteiro ML, Cabral GG, de Oliveira AL (2025) Pre-trained Code Language Models for Just-in-time Software Defect Prediction: An Empirical Study. In: 35th Brazilian Conference on Intelligent Systems, URL <https://bracis.sbc.org.br/2025/>, accepted for publication
- Ni C, Wang W, Yang K, Xia X, Liu K, Lo D (2022a) The best of both worlds: integrating semantic features with expert features for defect prediction and localization. In: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Association for Computing Machinery, New York, NY, USA, ESEC/FSE 2022, p 672–683, DOI 10.1145/3540250.3549165, URL <https://doi.org/10.1145/3540250.3549165>

- Ni C, Yang K, Xia X, Lo D, Chen X, Yang X (2022b) Defect Identification, Categorization, and Repair: Better Together. URL <https://arxiv.org/abs/2204.04856>, 2204.04856
- Niu F, Zhang E, Mayr-Dorn C, Assunção WKG, Huang L, Ge J, Luo B, Egyed A (2024) An extensive replication study of the ABLoTS approach for bug localization. *Empirical Software Engineering* 29(6):143, DOI 10.1007/s10664-024-10537-6, URL <https://doi.org/10.1007/s10664-024-10537-6>
- Pornprasit C, Tantithamthavorn CK (2021) JITLine: A Simpler, Better, Faster, Finer-grained Just-In-Time Defect Prediction. 2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR) pp 369–379, URL <https://api.semanticscholar.org/CorpusID:232223034>
- Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, KDD '16, p 1135–1144, DOI 10.1145/2939672.2939778, URL <https://doi.org/10.1145/2939672.2939778>
- Shehab MA, Khreich W, Hamou-Lhadj A, Sedki I (2024) Commit-time defect prediction using one-class classification. *Journal of Systems and Software* 208:111914, DOI <https://doi.org/10.1016/j.jss.2023.111914>, URL <https://www.sciencedirect.com/science/article/pii/S0164121223003096>
- Song L, Minku LL, Teng C, Yao X (2023) A Practical Human Labeling Method for Online Just-in-Time Software Defect Prediction. In: Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Association for Computing Machinery, New York, NY, USA, ESEC/FSE 2023, p 605–617, DOI 10.1145/3611643.3616307, URL <https://doi.org/10.1145/3611643.3616307>
- Subramanian VN (2024) BugLLM: Explainable Bug Localization through LLMs. Master's thesis, University of Waterloo, Canada, URL <hdl.handle.net/10012/21091>
- Tabassum S, Minku LL, Feng D (2023) Cross-Project Online Just-In-Time Software Defect Prediction. *IEEE Transactions on Software Engineering* 49(1):268–287, DOI 10.1109/TSE.2022.3150153
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, NIPS'17, p 6000–6010
- Wang X, Lu L, Yang Z, Tian Q, Lin H (2024) Parameter-Efficient Multi-classification Software Defect Detection Method Based on Pre-trained LLMs. *International Journal of Computational Intelligence Systems* 17(1):152, DOI 10.1007/s44196-024-00551-3
- Wu Q, Wang X, Wei D, Chen B, Dang Q (2025) Just-in-time software defect prediction method for non-stationary and imbalanced data streams. *Software Quality Journal* 33(1), DOI 10.1007/s11219-025-09711-w, URL

- <https://doi.org/10.1007/s11219-025-09711-w>, place: USA Publisher: Kluwer Academic Publishers
- Wu Y, Li Z, Zhang JM, Papadakis M, Harman M, Liu Y (2023) Large language models in fault localisation. URL <https://arxiv.org/abs/2308.15276>
- Zeng M, Wu Y, Ye Z, Xiong Y, Zhang X, Zhang L (2022) Fault localization via efficient probabilistic modeling of program semantics. In: Proceedings of the 44th International Conference on Software Engineering, Association for Computing Machinery, New York, NY, USA, ICSE '22, p 958–969, DOI 10.1145/3510003.3510073, URL <https://doi.org/10.1145/3510003.3510073>
- Zhou X, Xu B, Han D, Yang Z, He J, Lo D (2023) CCBERT: Self-Supervised Code Change Representation Learning. pp 182–193, DOI 10.1109/ICSME58846.2023.00028
- Zhou X, Han D, Lo D (2024) Bridging expert knowledge with deep learning techniques for just-in-time defect prediction. *Empirical Software Engineering* 30(1):37, DOI 10.1007/s10664-024-10591-0, URL <https://doi.org/10.1007/s10664-024-10591-0>
- Zhu X, Qiu T, Wang J, Lai X (2024) A novel instance-based method for cross-project just-in-time defect prediction. *Software: Practice and Experience* 54(6):1087–1117, DOI <https://doi.org/10.1002/spe.3316>, URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.3316>, <https://onlinelibrary.wiley.com/doi/pdf/10.1002/spe.3316>